AD_____

Award Number: DAMD17-97-1-7193

TITLE: Methods for Evaluating Mammography Imaging Techniques

PRINCIPAL INVESTIGATOR: Carolyn Rutter, Ph.D.

CONTRACTING ORGANIZATION: Center for Health Studies
Seattle, Washington 98101-1448

REPORT DATE: June 2000

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

NOTICE

USING GOVERNMENT DRAWINGS, SPECIFICATIONS, OR OTHER DATA INCLUDED IN THIS DOCUMENT FOR ANY PURPOSE OTHER THAN GOVERNMENT PROCUREMENT DOES NOT IN ANY WAY OBLIGATE THE U.S. GOVERNMENT. THE FACT THAT THE GOVERNMENT FORMULATED OR SUPPLIED THE DRAWINGS, SPECIFICATIONS, OR OTHER DATA DOES NOT LICENSE THE HOLDER OR ANY OTHER PERSON OR CORPORATION; OR CONVEY ANY RIGHTS OR PERMISSION TO MANUFACTURE, USE, OR SELL ANY PATENTED INVENTION THAT MAY RELATE TO THEM.

LIMITED RIGHTS LEGEND

Award Number: DAMD17-97-1-7193
Organization: Center for Health Studies
Location of Limited Rights Data (Pages):

Those portions of the technical data contained in this report marked as limited rights data shall not, without the written permission of the above contractor, be (a) released or disclosed outside the government, (b) used by the Government for manufacture or, in the case of computer software documentation, for preparing the same or similar computer software, or (c) used by a party other than the Government, except that the Government may release or disclose technical data to persons outside the Government, or permit the use of technical data by such persons, if (i) such release, disclosure, or use is necessary for emergency repair or overhaul or (ii) is a release or disclosure of technical data (other than detailed manufacturing or process data) to, or use of such data by, a foreign government that is in the interest of the Government and is required for evaluational or informational purposes, provided in either case that such release, disclosure or use is made subject to a prohibition that the person to whom the data is released or disclosed may not further use, release or disclose such data, and the contractor or subcontractor or subcontractor asserting the restriction is notified of such release, disclosure or use. This legend, together with the indications of the portions of this data which are subject to such limitations, shall be included on any reproduction hereof which includes any part of the portions subject to such limitations.

THIS TECHNICAL REPORT HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION.

_____ _____

_____ _____

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	June 2000	Annual (19 May 99 - 19 May 00)

4. TITLE AND SUBTITLE
Methods for Evaluating Mammography Imaging Techniques

5. FUNDING NUMBERS
DAMD17-97-1-7193

6. AUTHOR(S)
Carolyn Rutter, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Center for Health Studies
Seattle, Washington 98101-1448

E-MAIL:
rutter.c@ghc.org

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT: Distribution authorized to U.S. Government agencies only (proprietary information, Jun 00). Other requests for this document shall be referred to U.S. Army Medical Research and Materiel Command, 504 Scott Street, Fort Detrick, Maryland 21702-5012.

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

The purpose of this award is to enable Dr. Rutter to develop biostatistical methods for evaluating the accuracy of breast cancer screening. This four year program includes advanced training in the epidemiology of breast cancer, training in clinical detection of breast cancer, development of statistical methodology, and graduate teaching. A basic knowledge of the epidemiology, disease process and detection of breast cancer guides the development of statistical methods.

During this third funding year, Dr. Rutter has continued to expand her knowledge of breast cancer epidemiology and detection. She has published two articles during the third funding year. The 1st describes bootstrap estimation of accuracy statistics when patients are assessed at multiple patient sites. The 2nd published article compares performance of mammographers in a test setting to performance in clinical practice. Dr. Rutter also has a third article that is under review by JAMA that compares changes in breast density among women who initiate, discontinue, and continue use of hormone replacement therapy.

During her fourth funding year, Dr. Rutter will teach an introductory graduate level statistics course and will focus on methods for estimating sensitivity and specificity that incorporate growth curve models.

14. SUBJECT TERMS
Breast Cancer

15. NUMBER OF PAGES
63

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

___ Where copyrighted material is quoted, permission has been obtained to use such material.

___ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

___ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

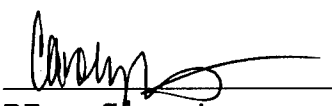
N/A In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

X For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

N/A In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

N/A In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

N/A In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.



6-15-00

PI - Signature Date

Table of Contents

Introduction

The purpose of this Department of Defense Breast Cancer Research Program Career Development Award is to enable Dr. Rutter to develop biostatistical methods for evaluating the accuracy of breast cancer screening. This four year program includes advanced training in the epidemiology of breast cancer, training in clinical detection of breast cancer, development of statistical methodology, and graduate teaching. A basic knowledge of the epidemiology, disease process and detection of breast cancer will guide the development of statistical methods. During the third funding year, Dr. Rutter's research has shifted away from ordinal measures based on ROC analyses and towards dichotomous outcomes (see Technical Objective 3). There has also been a shift in emphasis, away from purely statistical research and toward epidemiological and health services research. During her fourth funding year, Dr. Rutter will teach an introductory graduate level statistics course (see Technical Objective 5), and will focus on methods for estimating sensitivity and specificity that incorporate growth curve models (see Technical Objective 4).

Achievement of Year 3 Technical Objectives

Technical Objective 1: Gain additional training in breast cancer epidemiology, detection and treatment.

Dr. Rutter has continued to expand her knowledge about breast cancer through attendance scientific seminars at the Fred Hutchinson Cancer Research Center (FHCRC) and the University of Washington (UW). Dr. Rutter also participates in a Diagnostic Methods working group that includes faculty from both FHCRC and UW. Dr. Rutter also attends Breast Cancer Surveillance Consortium (BCSC) meetings that have provided her with important practical information about radiologists' interpretation of mammograms, and the timing and execution of diagnostic procedures. Through additional reading and analysis of BCSC data, Dr. Rutter has gained specialized knowledge about the interrelationships among hormone replacement therapy, breast density and breast cancer.

Technical Objective 2: Develop methods for multiple patient assessments.

This objective was completed during year two of this CDA award. The article describing nonparametric bootstrap ROC estimation for correlated data is currently in press at *Academic Radiology* (see Appendix B). This article compares an iterative bootstrap estimation approach to estimating and comparing the area under the ROC curve with a non-iterative method that uses sums of squares to adjust variance estimates for correlation between observations.[1] Both methods are theoretically valid, and both perform well in a simple situation. However, the bootstrap estimator can more easily be used in complex sampling situations that include multiple sources of correlation.

Technical Objective 3: Extend exact methods for ordinal regression models.

This is no longer a key research objective for two reasons. First, ordinal regression models are of limited usefulness in the mammographic screening setting. Instead, most interest focuses on the sensitivity and specificity of mammography based on particular definitions of positivity. In addition, most mammograms are assessed using BI-RADS ratings[2], and these ratings do not use a pure ordinal scale. The second reason for moving away from this objective is that Dr. Rutter is using data from the BCSC, a surveillance project that captures information from large populations of women.[3] Although small sample problems are of interest when diagnostic tests are compared using small samples, they are less useful in the context of evaluation of screening tests based on large population-based samples.

Technical Objective 4: Develop methods to adjust for error in measurement of disease status.

Dr. Rutter is currently working to develop new methods that adjust for error in measurement of disease status. The proposed work will use hierarchical Bayesian modelling approaches to incorporate models for tumor growth rate into estimation of sensitivity and specificity. Currently, a fixed one or two year follow-up period is used to define true state. A fixed follow-up period approach to defining disease outcome ignores variability in growth rates by age and tumor type. There is some evidence that tumors in younger women grow faster than tumors in older women [4-6]. Growth rates may also vary by tumor type. Thus, sensitivity and specificity are incorrectly estimated for young women or women with particular tumor types. Such biased estimates of sensitivity and specificity of screening mammography ultimately affect health policy decisions.

During the fourth funding year (and potentially extending into a fifth year based on a no-cost extension), Dr. Rutter will examine estimation of sensitivity and specificity using probabilistic disease estimates. That is, rather than estimating a disease state (present/absent), models will estimate the probability of screen detectable disease at the time of mammography. Modesl for the probability of screen detectable disease will use state of the art models for tumor growth rate [5,7-12], tumor size for patients diagnosed with cancer, and the time from screening mammography to either biopsy or end of follow-up.

Because these models focus on screen detectability, preliminary research will estimate the minimum size of screen detectable tumors using BCSC data. In particular, we will examine the distribution of tumor size by age and breast density among screen-detected cases, assessing the impact of these covariates on the distribution of screen detectable tumor size. As part of these analyses, we will examine the distribution of tumor size by age and breast density among clinically detected cases. Publication of these results will provide important information for research examing breast cancer modelling.

Primary analyses will examine the effects of incorporating tumor growth rates on estimates of sensitivity and specificity. These analyses will explore the impact of various growth rate models on accuracy estimates, as well as the impact of alternate definitions of screen detectable tumor size, focusing on definitions that depend on breast density.

Subsequent analyses will focus on improvement of growth models based on observed mammographic tumor size. Challenges faced in this research include appropriate treatment of true interval cancers (i.e., cancers that are not observable at the mammogram prior to clinical discovery) and treatment of calcifications that have no measurable tumor mass. Because these analyses require primary data collection from mammograms, they are beyond the scope of the current proposed research, and instead will be the subject of future grant applications.

Technical Objective 5: Develop and teach a course in methods for assessing diagnostic tests.

This technical objective is no longer possible, as Dr. Margaret Pepe, a full professor at University of Washington, will teach a special topics course on Medical Diagnostic Testing in Spring 2000. Instead, Dr. Rutter will gain teaching experience by teaching introductory biostatistics to health services students (Biostatistics 509). While teaching a graduate level survey course was not the original intent of the training grant, this experience will be valuable training and will enhance Dr. Rutter's career development. Successfully teaching this introductory course will increase Dr. Rutter's chance of teaching special topics courses in the future.

Key Research Accomplishments

During this third funding year, Dr. Rutter has continued to expand her knowledge of breast cancer epidemiology and detection. She has published two articles during the third funding year. The first [13] describes bootstrap estimation of accuracy statistics (TP, FP, AUC) when patients are assessed at multiple patient sites. The second published article [14] compares performance of mammographers in a test setting to performance in clinical practice. Dr. Rutter also has a third article that is under review by JAMA (see Appendix D) that compares changes in breast density among women who initiate, discontinue, and continue use of hormone replacement therapy.

Reportable Outcomes

1. Rutter CM, Taplin S. "Assessing Mammographers' Accuracy: A comparison of clinical and test performance," *Journal of Clinical Epidemiology,* 2000: 443-450.
2. Rutter CM. Bootstrap estimation of diagnostic accuracy using patient-clustered data. in review, *in press, Academic Radiology.*
3. Taplin S, Rutter CM, Elmore JG, Seger D, White E, Brenner RJ. "Accuracy of Screening Mammography on Single Versus Independent Double Reading," *American Journal of Roentgenology,* 2000 174:1257-62.
4. Rutter CM, Mandelson MT, Laya MB, Seger DJ, Taplin S. "Changes in Breast Density associated with Initiation, Discontinuation, and Continuing use of Hormone Replacement Therapy (HRT),"submitted to *JAMA.*
5. Rutter CM, Gatsonis CG. A hierarchical regression approach to meta-analysis of diagnostic test accuracy. in submission *Statistics in Medicine,* in process of revising for resubmission.
6. Participated in Workshop on Assessment and Improvement of Interpretive Skills in Mammography, sponsored by the American Cancer Society, June 11/12, 1999.
7. Presented " Assessing mammographers' accuracy: A comparison of clinical and test performance" at the International Conference on Health Policy Statistics: Methodologic Issues in Health Services and Outcomes Research, Santa Monica, California, December 3-5, 1999

Conclusions

- Regarding bootstrap estimation: Bootstrap estimation of the area under the receiver operating characteristic curve, sensitivity, and specificity allows simple and accurate calculation of confidence intervals for single tests and comparisons between tests.
- Regarding the use of tests data sets: Direct estimation of mammographer's clinical accuracy requires the ability to capture screening assessments and correctly identify which screened women have breast cancer. Use of screening sets offers an attractive alternative method for estimating mammographers' accuracy. Unfortunately, we found that there was little concordance between performance on a test film set and performance in clinical practice. There is the potential for bias in both types of assessment, and our research cannot distinguish which approach is best. It does, however, raises questions about construction of and use of test film sets.
- Regarding use of HRT and breast density: Initiation of hormone replacement therapy (HRT) has been shown to increase breast density [15-19]. Several lines of evidence indicate that breast density is strongly related to breast cancer risk [20-23] and that increased density decreases mammographic sensitivity [24]. Using an cohort of 5213 naturally postmenopausal women 40 to 96 years old, we used consecutive mammograms and pharmacy records to examine the relationship between initiation, cessation and continuing use of HRT on breast density. We found that women who initiated HRT were more likely than nonusers to show increases in density (OR=3.24, 95% CI (2.47,4.23)), while women

who discontinued were more likely show decreases in density (OR=1.92, 95% CI (1.03,3.35)), and women who continued use of HRT were more likely to show both increases in density (OR=1.37, 95% CI (0.89,2.06)) and sustained high density (OR=1.72, 95% CI (1.50,1.98)). Continuing HRT use was more strongly associated with sustained high density among women with high BMI ($p<0.05$). These results provide strong evidence that breast density changes associated with HRT are dynamic, increasing with initiation and decreasing with discontinuation. Continued HRT use results in persistent changes, particularly among women with high BMI.

- Regarding estimation of sensitivity and specificity for screening mammography: Several lines of evidence demonstrate variability in tumor growth rates. Estimates of mammography performance can be improved through incorporation of growth rate model.

References

1. Obuchowski NA. "Nonparametric Analysis of Clustered ROC Curve Data," *Biometrics,* 53: 567-578, 1997.
2. Breast Imaging Reporting and Data System (BI-RADS), Third Edition, American College of Radiolgy : Reston, VA, 1998
3. Ballard-Barbash R, Taplin SH, Yankaskas BC, et al. "Breast Cancer Surveillance Consortium: A National Mammography Screening and Outcomes Database," *AJR,* 169:1001-8, 1997.
4. Brekelmans CT, v Gorp JM, Peeters PH, Collette HJ. "Histopathology and growth rate of interval breast carcinoma. Characterization of different subgroups," *Cancer,* 71:1220-8, 1996.
5. Peer PGM, v Dijck JAA, Hendriks JHCL, Holland R, Verbeek ALM. "Age-Dependent Growth Rate of Primary Breast Cancer," *Cancer,* 71:3547-51, 1993.
6. Breckelmans CT, Westers P, Faber JA, Peeters PH, Collette HJ, "Age Specific Sensitivity and Sojourn Time in a Breast Cancer Screening Programme (DOM) in the Netherlands: A Comparison of Different Methods," *J Epidemiol Community Health,* 50:68-71, 1996.
7. v Fournier D, Webber E, Hoeffken W, Bauer M, Kubli F, Barth V. "Growth Rate of 147 Mammary Carcinomas," *Cancer,* 45:2198-207, 1980.
8. Spratt JA, v Fournier D, Spratt JS, Weber EE. "Decelerating Growth and Human Breast Cancer," *Cancer,* 171:2013-9, 1993.
9. Spratt JA, v Fournier D, Spratt JS, Weber EE. "Mammographic Assessment of Human Breast Cancer Growth and Duratin," *Cancer,* 171:2020-6, 1993.
10. Hart D, Shochat E, Agur Z. "The Growth Law of Primary Breast Cancer as Inferred from Mammography Screening Trials Data," *Br J Cancer,* 78:382-8, 1998.
11. Parmigiani G. "Decision Models in Screening for Breast Cancer," *Bayesian Statistics* Bernadno JM, Berger JO, Dawid AP, Smith AFM (Eds), New York: Oxford Univeristy Press, 1999.
12. Ashih HW, Berry DA, Parmigianni G. "Modeling Natural History of Breast Cancer Tumor Growth," *unpublished technical report.*
13. Rutter CM, "Bootstrap Estimation of Diagnostic Accuracy using Patient-clustered Data," in press, *Academic Radiology.*
14. Rutter CM, Taplin S. "Assessing Mammographers' Accuracy: A comparison of clinical and test performance," *Journal of Clinical Epidemiology,* 2000: 443-450.
15. Stomper PC, Van Voorhis BJ, Ravnikar VA, Meyer JE. Mammographic changes associated with postmenopausal hormone replacement therapy: a longitudinal study. *Radiolgoy.* 1990;174:487-90.
16. Kaufman Z, Garstin WI, Hays R, et al. The mammographic parencymal patterns of women on hormonal replacement therapy. *Clinical Radiology.* 1991;43:389-92.

17. Laya MB, Gallagher JC, Schreiman JS, et al. Effect of postmenopausal hormone replacement therapy on mammographic density and parenchymal pattern. *Radiology.* 1995;196: 433-7.
18. Greendale GA, Reboussin BA, Sie A, et al. Effects of Estrogen and Estrogen-Progestin on Mammographic Parenchymal Density. *Annals of Internal Medicine.* 1999;130: 262-269.
19. Lundstrom E, Wilezek B, von Palffy Z, et al. Mammographic breast density during hormone replacement therapy: Differences according to treatment. *American Journal of Obstetrics and Gynecology.* 1999;18:348-352.
20. Litherland JC, Stallard S, Hole D, Cordiner C. The effect of hormone replacement therapy on the sensitivity of screening mammograms. *Clin Radiol.* 1999;54:285-8.
21. Seradour B, Esteve J, Heid P, Jacquemier J. Hormone replacement therapy and screening mammography: analysis of the results in the Bouches du Rhone programme. *J Med Screen.* 1999;6:99-102.
22. Kavanagh AM, Mitchell H, Giles GG. Hormone replacement therapy and accuracy of mammographic screening. *Lancet.* 2000;355:270-274.
23. Saftlas AF, Szklo M. Mammographic parenchymal patterns and breast cancer risk. *Epidemiol Review.* 1987;9:146-74.
24. Mandelson MT, Oestreicher N, Porter PL, Taplin SH, White E. Breast density as a predictor of mammographic detection: Comparison of interval- and screen-detected cancers. *JNCI. in press.*

Appendices

Appendix A. Statement of Work

Technical Objective 1: Gain additional training in breast cancer epidemiology, detection and treatment.

Task 1: Months 1-4: Review of information on the epidemiology, diagnosis and treatment of breast cancer as suggested by Dr. Margaret Mandelson.

Task 2: Months 1-48: Attend seminars sponsored by the Seattle Breast Cancer Research Program.

Technical Objective 2: Statistical research, aim 1: develop methods for multiple patient assessments.

Task 3: Month 6: Review current research for generalized estimating equation and random effect approaches for nonlinear models.

Task 4: Months -11: Test bootstrap, robust covariance adjustment and generalized estimating equation methods for breast-level analyses using simulation studies.

Task 5: Months 12-21: Develop methods for woman-level analysis, possibly including software development for random effects in generalized ordinal regression models.

Technical Objective 3: Statistical research, aim 2: extend exact methods for ordinal regression models

Task 6: Month 22: Review current research in exact methods.

Task 7: Months 23-34: Extend exact methods and write computational algorithms and programs to compute distributions of sufficient statistics.

Technical Objective 4: Statistical research, aim 3: Develop methods to adjust for measurement error in disease status

Task 8: Month 36: Review current research in errors-in-measurement models.

Task 9: Months 37-48: Develop simple combined corrections for verification and follow-up bias. These methods will be extended to allow adjustments in general ordinal regression models.

Technical Objective 5: Develop and teach a course in methods for assessing diagnostic tests.

Task 10: Months 1-24: Collect relevant references and outlining lectures for the methods course. During this time, specific lectures may be presented in other University of Washington courses.

Task 11: Months 25-36: Offer methods course at University of Washington through the Department of Biostatistics.

Appendix B:
Bootstrap Estimation of Diagnostic Accuracy with Patient-Clustered Data

running head: BOOTSTRAP ESTIMATION OF DIAGNOSTIC ACCURACY

Original Investigations

Bootstrap Estimation of Diagnostic Accuracy with Patient-clustered Data[1]

Carolyn M. Rutter, PhD

Rationale and Objectives. The purpose of this study was to describe a simple bootstrap approach for estimating sensitivity, specificity, and the area under the receiver operating characteristic curve for multisite test outcome data.<a>

Materials and Methods. The performance of bootstrap estimates was evaluated and compared with that of analytic estimates ~~by~~ using a simulation study. Bootstrapping was demonstrated ~~by~~ using data from a previous study comparing two angiographic methods.

Results. Analytic and bootstrap estimates had similar coverage rates for 95% confidence intervals. With many sites per patient, bootstrap estimates had slightly better coverage than analytic estimates. Bootstrap percentile intervals had better coverage than asymptotic normal bootstrap intervals.

Conclusion. Bootstrapping is a useful method for estimating confidence intervals for the area under the receiver operating characteristic curve, sensitivity, and specificity when data are

---

[a] Au: Please note journal style calls for the purpose as stated in the abstract to match that as stated in the text. Please confirm all necessary information has been included. ✓

correlated.

Key Words. Area under the receiver-operating characteristic curve (AUC); sensitivity; specificity.

Acad Radiol 2000; 7:000-000

[1]From the Group Health Cooperative of Puget Sound, Center for Health Studies, 1730 Minor Ave, Suite 1600, Seattle, WA 98101.

Received July 6, 1999; revision requested December 2; revision received December 15; accepted December 23. **Address correspondence to** C.M.R.

Diagnostic evaluation often requires simultaneous assessment of disease at multiple body sites. Examples of multisite diagnostic assessments include screening mammography to detect breast cancer, computed tomography (CT) of the liver to detect metastatic colorectal cancer (1), and magnetic resonance (MR) angiography of leg vessels to detect occlusive peripheral vascular disease (2). Although the accuracy of these multisite tests can be estimated by using information from a single body site, studies using all available information have greater statistical power. Reducing the site-level data to patient-level data is the simplest approach to multisite diagnostic assessment. Composite patient-level measures of true state and test outcome, however, reduce the amount of information regarding test accuracy

contained in multisite assessments. These composite measures also ignore disease localization, which can be more important than global determination of disease presence when making treatment decisions.

Estimates of diagnostic accuracy that use multisite data must account for within-patient correlation. Methods of handling multiple assessment of a single site, by using different modalities or readers, are well developed. Song (3) provides an overview of current approaches. These methods require that patients are either diseased or not diseased, and they can be used to evaluate multisite assessments when the true state is constant across sites within patients.

When both disease state and test outcomes are dichotomous, marginal regression models can be used to estimate the sensitivity, the specificity, and the effects of patient covariates on the sensitivity and specificity (4). This flexible modeling approach provides standard errors that reflect the clustering of data within patients because of the assessment of multiple sites and assessment by multiple readers or modalities.

When the disease state is dichotomous, logistic regression models can be used to estimate the relationship between true state and test outcomes (2). When data are clustered within patients, standard methods can be used to adjust the logistic regression coefficient covariance matrix for within-patient

correlation (5). The logistic model conditions on the test results and estimates their association with disease state.<b> These models do not result in standard measures of accuracy, however, thereby making comparisons with the results of other studies difficult.

Obuchowski (6) described a method for estimating ~~to estimate~~ standard errors for the area under the empirical receiver operating characteristic curve (AUC) on the basis of the sums of squares. This method allows estimation of the standard error of the AUC for a single test or of the difference between AUC statistics for two tests. Obuchowski's approach requires the definition and calculation of appropriate sums of squares, however, and this can become complicated with multiple sources of correlation (eg, when patients are evaluated at multiple sites by more than one test and each test is independently evaluated by more than one reader).

Pepe (7) proposed a general regression method that allows multisite assessments. This regression approach estimates the effects of covariates on the receiver-operating characteristic (ROC) curve. Interpretation of the regression coefficients depends on the functional form that is chosen for the ROC curve. Coefficients estimated from a logistic model can be interpreted as the log-odds of correctly classifying a diseased subject for a

---

[b] Au: Please confirm/clarify: "conditions on" correct? yes

fixed specificity. Pepe suggests using bootstrap resampling to estimate standard errors of regression coefficients when correlated data are included in these models.

This study demonstrates a simple bootstrap approach for estimating sensitivity, specificity, and AUC for multisite test outcome data. This bootstrap approach is useful for simple comparisons between tests in situations with no covariates. When regression approaches are used, bootstrap estimates can provide supplemental descriptive statistics. This approach also is easy to use with multiple sources of correlation, and the resulting confidence intervals (CIs) are asymptotically consistent.

MATERIALS AND METHODS[c]

Nonparametric Accuracy Statistics

The accuracy of an imaging examination depends on the radiologist's interpretation of disease state. These interpretations typically are measured by using a five-point ordinal scale that ranges from "definitely not diseased" to "definitely diseased." Sensitivity, specificity, and the AUC are the basic statistics used to measure test accuracy. These statistics condition on the true disease state, treating it as being fixed and known and treating test outcomes (ie, ratings) as being randomly distributed. When the disease state is known

---

[c] Au: Please note text heads have been edited to conform with journal style. Please confirm heads are accurate and as meant. ~~ok. another possible place for this could come before "Angiography Study" Since the sections "Nonparametric Accuracy Statistics" & "Bootstrap Estimates" are background.~~ ok. this is the best place.

without error, these accuracy statistics are independent of disease prevalence.

When test outcomes are dichotomous, sensitivity and specificity measure test accuracy. Sensitivity is the probability of a positive test outcome (ie, indicating presence of disease) when the target disease is present. Specificity is the probability of a negative test outcome when the target disease is absent. When test outcomes are ordinal, sensitivity and specificity can be calculated by dichotomizing the outcomes. A single sensitivity-specificity pair, however, cannot completely describe the accuracy of an ordinal test, because both rates depend on test stringency. Analysis of the ROC curve accounts for the tradeoff in these rates as the test stringency varies. For example, suppose that the ordinal outcome of a diagnostic test t_i takes values in $\{1, 2, \ldots, K\}$, with increasing values of t_i corresponding to stronger evidence of disease. There are $K + 1$ possible ways to dichotomize the ordinal test, including "all positive" and "none positive," and each way is associated with a sensitivity-specificity pair.

The empirical ROC curve is drawn by plotting pairs of observed rates, (1 - specificity) versus sensitivity, and connecting the $K + 1$ consecutive points with straight lines. The empirical ROC curve provides a simple, graphical description of test performance.

The overall accuracy of an ordinal test can be summarized by the AUC, which estimates the probability of correctly ranking a randomly selected (diseased, not-diseased) pair on the ordinal test scale. The AUC ranges from 0 to 1, with the value 1 corresponding to a perfect diagnostic test. A test that is no better than chance has an AUC equal to one-half. The AUC statistic is unbiased, and it is asymptotically and normally distributed. The test of H_o:AUC = ½ based on the asymptotic distribution is equivalent to a Mann-Whitney test (8). Essentially, the AUC test is a test for differences in the distribution of test outcomes among the diseased and not-diseased groups.

Bootstrap Estimation

Sensitivity, specificity, and the AUC are all generalized U-statistics of order 1, and each statistic is a sum of the functions of statistically independent quantities (9). Because sensitivity, specificity, and the AUC are U-statistics, bootstrap resampling provides consistent point and interval estimates (10,11).

Let $\mathbf{t}_i = (t_{i1}, t_{i2}, \ldots, t_{im})'$ be the vector of ordinal test outcomes across m sites for the ith subject, and let $\mathbf{d}_i = (d_{i1}, d_{i2}, \ldots, d_{im})'$ be the corresponding vector of true states, where $d_{ij} = 1$ if the jth site of the ith patient is diseased and $d_{ij} = 0$ if otherwise. Written in U-statistic form, sensitivity

and specificity for the kth cut point are

[COMP: PLEASE PICKUP EQUATION E1]

and

[COMP: PLEASE PICKUP EQUATION E2]

with kernel function $\phi_k(t_i, d_i) = \sum_j \delta_k(t_{ij}) d_{ij}$, where $\delta_k(t) = 1$ if $t \geq k$ and $\delta_k(t) = 0$ if otherwise. The associated sample sizes are $n_D = \sum_i \sum_j d_{ij}$ and $n_{\bar{D}} = \sum_i \sum_j (1 - d_{ij})$. Here, D indicates the presence of disease and $\bar{D}$ indicates the absence of disease.

The AUC statistic is given by

[COMP: PLEASE PICKUP EQUATION E3]

with kernel function

[COMP: PLEASE PICKUP EQUATIONS E4]

When both diseased and not-diseased sites can occur within a patient, the sum corresponding to the AUC statistic includes the functions of correlated pairs of diseased and not-diseased

observations, thereby violating the properties of U-statistics. Relatively few correlated (D, $\overline{D}$) pairs, however, are included in the sum. Let p_p be the patient-prevalence of disease, and let p_s be the expected proportion of sites with disease given that a patient has disease. If all patients with disease have the same number of affected sites, the proportion of correlated (D, $\overline{D}$) pairs is

[COMP: PLEASE PICKUP EQUATION E5]

When all patients have a single disease state, $p_s = 1$, there are no correlated (D, $\overline{D}$) pairs. The number of correlated pairs is maximized at $1/N$ when all N patients have disease ($p_p = 1$). With correlated pairs, the U-statistic properties of the AUC statistic can be maintained by excluding correlated pairs from the AUC sums. In most cases, this exclusion is unnecessary because the number of correlated comparisons quickly becomes negligible as the sample size increases. This study, however, examines bootstrap resampling applied directly to AUC statistics without excluding correlated pairs.

Bootstrap samples were constructed by stratifying patients on overall disease state (any or none) and then drawing patients (ie, the independent units) with replacement from these strata. Resampling patient-level data incorporates all sources of within-

patient variability. Stratifying the bootstrap samples by patient-level disease state corresponds to conditioning on true disease state. Disease-state stratification was used to ensure that all the accuracy statistics examined (ie, sensitivity, specificity, AUC) were estimable. Because each of these statistics conditions on disease state, stratified sampling does not bias the point estimates. Accuracy statistics were calculated for each bootstrap sample, and expected accuracy was estimated by using the average of each statistic across the bootstrap samples. The accuracy of two tests can be compared by calculating the difference in the accuracy statistics for each bootstrap sample and incorporating between-test correlation. Standard errors were estimated ~~by~~ using the observed standard errors across the bootstrap samples; standard error estimates should be based on at least 100 draws. CIs can be estimated by using the bootstrap-estimated standard errors with a normal approximation. CIs also can be estimated by using percentiles, although this requires at least 1,000 bootstrap draws.

Angiographic Study

~~Conventional~~ Contrast angiography is the usual method of mapping vascular occlusion before coronary artery bypass graft surgery in patients with peripheral vascular disease. MR angiography is an alternative method of obtaining the same diagnostic information. MR angiography is less invasive than conventional angiography,

however, because it does not require injection of contrast material. The ability of both conventional angiography and MR angiography to identify open vessel segments correctly was compared by using a prospective study, with intraoperative angiography considered to be the gold standard (2). Analyses were based on 96 patients with peripheral vascular disease for whom intraoperative angiographic results and at least one preoperative angiographic test result were available. Eleven of these patients did not undergo MR angiography.

Patients were evaluated at 15 sites (ie, vessel segments). On average, 33% of each patient's vessel segments were occluded. Overall, 335 (36%) of 932 segments with gold standard information were occluded. Study radiologists rated the occlusion of each vessel segments by using a five-point scale: (*a*) normal, (*b*) minimal disease (ie, <50% stenosis), (*c*) stenotic (ie, a single lesion with ≥50% stenosis but not fully occluded), (*d*) diffuse disease (multiple lesions with ≥50% stenosis but not fully occluded); and (*e*) fully occluded. Ratings within patients were moderately correlated, with similar degrees of correlation for the two imaging examinations. Overall, correlation (based on Kendall τ) was 0.20 for conventional angiography and 0.19 for MR angiography. Correlation between sites with the same disease state was 0.49 for conventional angiography and 0.46 for MR angiography. Correlation between sites with different disease

states was -0.34 for conventional angiography and -0.36 for MR angiography. Correlation between conventional angiography and MR angiographic ratings of the same site was 0.62.

Original analyses examined both detection of near-normal (ie, patent) vessel segments (ratings 1 and 2) and detection of open segments (ratings 1-4). CT and MR angiography had similar accuracy when identifying open vessel segments. Both modalities had an 81% specificity, with conventional angiography having an 83% sensitivity and MR angiography an 85% sensitivity. When identifying patent segments, conventional angiography was less sensitive than MR angiography (77% vs 82%) but also more specific (92% vs 84%). On the basis of these descriptive data and statistical tests for the differences in odds ratios, the original investigators concluded that conventional angiography and MR angiography had similar diagnostic abilities.

Bootstrap estimation allowed us to estimate AUC statistics for patent segments, to examine whether differences likely resulted from a threshold effect, and to place CIs on estimated sensitivity and specificity. Bootstrap percentile intervals were based on 1,000 bootstrap samples. The bootstrap estimate for the sensitivity of conventional angiography was 76% (95% CI: 70.5, 81.8). For MR angiography, the bootstrap estimate for sensitivity was 82% (95% CI: 76.8, 87.0). Bootstrap estimates of specificity were 93% (95% CI: 89.8, 95.9) for conventional angiography and

84% (95% CI: 79.4, 88.1) for MR angiography.

Both conventional angiography and MR angiography had similar empirical AUC statistics. For conventional angiography, the empirical AUC was 0.879 (95% CI: 0.847, 0.910). For MR angiography, the empirical AUC was 0.874 (95% CI: 0.844, 0.904). The bootstrap estimate of the difference in AUC statistics was 0.005 (95% CI: -0.035, 0.044).

Simulation Study

This simulation study describes characteristics of bootstrap accuracy and compares them to the analytic estimates previously reported by Obuchowski (6). Bootstrap CIs determined on the basis of normal approximations used 100 bootstrap samples. Bootstrap CIs determined on the basis of percentiles used 1,000 bootstrap samples. Comparisons focused on the observed coverage of 95% CIs for the differences between two AUC statistics. The description of bootstrap estimates also included coverage rates for the estimated specificity.

Simulated data represent comparisons between two tests (A and B), with outcomes being scored with a five-point ordinal scale. Test A has an empirical AUC of 0.8 and specificities of 0.5, 0.7, 0.9, and 0.95. Test B had the same specificities and an AUC statistic of 0.80 or 0.85. The ability of the bootstrap to handle multiple sources of variability was evaluated by simulating the outcomes for two readers per test. The overall diagnostic

accuracy of each test was determined on the basis of the average of the two readers' AUC statistics. Data simulated for two readers assumed that readers evaluating the same test had equal accuracy, with the same specificities and the same AUC statistics. Two-reader bootstrap AUC estimates were calculated by estimating each reader's AUC statistic and then averaging these values within each bootstrap sample.

Ordinal test outcomes were simulated by categorizing continuous multivariate normal (MVN) pseudodeviates. One MVN pseudodeviate of length $4m$ was generated for each patient-observation, with m being the number of sites within patients.[d] Each independent MVN pseudodeviate represented a single patient's unobservable, continuous test outcome for two tests and two readers. Within-patient correlation was induced on the continuous scale. The simulations examined the characteristics of estimators for three within-subject correlation structures: (*a*) independent, (*b*) compound symmetry, and (*c*) disease-dependent. Under the compound symmetry structure, multiple observations within subjects were equicorrelated (correlation, 0.50). The disease-dependent structure was identical to the compound symmetry structure with one exception: Under the disease-dependent structure, observations from sites with different disease states

---

[d] Au: Do you mean the number of sites within each patient or overall (for the patient group)?

(ie, [D, $\bar{D}$] pairs) were negatively correlated (correlation, -0.50).

The simulation examined three sampling scenarios. In the first scenario (ie, small N), 100 patients (50 with disease and 50 without) were evaluated at four sites. In the second scenario (ie, large m), 100 patients (all with disease) were evaluated at 15 sites. In the third scenario (ie, large N), 500 patients (250 with disease and 250 without) were evaluated at four sites. For each scenario, patients with disease were expected to have disease at half the sites examined. The number of disease-positive sites for each patient was simulated by using a binomial random-number generator. Ordinal ratings were derived from MVN deviates by assuming an underlying bivariate, normal ROC model (12). In other words, "cut points" for each of the five rating categories were set equal to $\theta_0 = -\infty$; $\theta_k = \Phi^{-1}(1 - \text{specificity}_k)$; $k = 1,\ldots, 4$; and $\theta_5 = +\infty$. Given μ and θ, the sensitivities were $\text{sensitivity}_k = \Phi(\theta_k + \mu)$. The desired empirical AUC statistics were obtained with $\mu = 1.29$ for an AUC of 0.80 and $\mu = 1.949$ for an AUC of 0.85. For disease-negative sites, the ordinal rating corresponding to the MVN deviate y was equal to k when $\theta_{k-1} < y < \theta_k$. For disease-positive sites, the MVN deviates were first shifted by an appropriate μ, with simulated ratings determined on the basis of categorizing $y + \mu$.

Simulation results were determined on the basis of 5,000

simulated data sets for each combination of AUC_B (0.80 or 0.85), <e> sampling scenario (small *N*, large *m*, or large *N*), and correlation structure (independent, equicorrelated, and disease-dependent).

RESULTS

Table 1 shows the observed within-patient correlations for the simulated categoric data. These rating data are inherently correlated, because diseased sites are more likely than not-diseased sites have high scores.

Table 2 shows coverage rates of the 95% CIs for the difference between AUC_A and AUC_B as determined on the basis of Obuchowski's analytic estimator, the single-reader bootstrap percentile interval, and the two-reader bootstrap percentile interval. Coverage rates for normal-approximation bootstrap intervals are not shown, because they were similar to those of the percentile intervals but with slightly poorer coverage properties. In general, coverage rates of the normal-approximation bootstrap interval fell between the coverage rates for analytic and bootstrap percentile intervals. Table 2 lists coverage rates by sampling scenario and correlation structure, because true differences between the two AUC statistics did not affect the coverage. The bootstrap and analytic CIs had very similar coverage rates for the small *N* and the large *N* scenarios.

---

[e] Au: Any mention of AUC_A needed here?
Could add in parens that $AUC_A = 0.8$ for all scenarios - this is mentioned @ bottom of p13.

Bootstrap intervals had better coverage for the large *m* scenario. ~~Both~~ Single- and two-reader bootstrap intervals had similar rates of coverage.

~~Both~~ The analytic and the single-reader bootstrap estimates had a similar mean squared error. Across the simulated data sets, the mean squared error of the bootstrap estimate for one reader was less than 0.1% higher than the mean squared error for the analytic estimate. The mean squared error for the two-reader bootstrap estimates was approximately half the mean squared error of either single-reader estimate.

Table 3 shows the coverage rates of bootstrap percentile interval estimates by specificity. Generally, coverage rates were less than the nominal level but improved as the specificity decreased from 0.95 to 0.50 and as the amount of data available for estimation increased. Percentile intervals had better coverage rates than the asymptotic normal intervals (not shown). When the specificity was 0.95, a few (≈0.5%)<f> of the asymptotic normal bootstrap intervals fell outside of the (0, 1) range.

DISCUSSION

Diagnostic evaluation often involves testing patients at multiple sites. Bootstrap and analytic estimation methods allow for simple comparisons of AUC statistics (paired) on ~~the basis of~~

---

[f] Au: Please provide raw numbers to accompany this percentage. 140/30,000

clustered patient data. These methods are asymptotically consistent; however, diagnostic tests rarely are evaluated on the basis of large samples. We used a simulation study to evaluate the small-sample characteristics of Obuchowski's analytic AUC estimator and bootstrap AUC estimators applied to ordinal test data. When comparing two tests with one reader per test, the bootstrap and analytic estimators had similar performance. Both methods produced CIs ~~with~~ with observed coverage rates that were below the nominal level. Coverage rates of bootstrap percentile CIs were nearly identical with the asymptotic normal intervals for AUC statistics. Percentile intervals, however, had better coverage than asymptotic normal intervals for proportions. These methods are asymptotically consistent, but results of the simulations suggest that when the test outcomes are ordinal and the tests themselves are relatively accurate, large samples are needed before asymptotic results hold.

The simulations in this study demonstrated poorer performance for Obuchowski's estimator than was originally reported. There are important differences ~~were found~~ between the simulations in this study and those reported by Obuchowski. Two key differences are the site-level prevalence of disease and the scale of the test outcome. In the small-sample setting, the patient-level prevalence of disease was 50%, and among patients with disease, an average of 50% of sites were affected, thereby

resulting in an overall site-level prevalence of 25%. Obuchowski simulated data with an overall site-level prevalence of 50%. Obuchowski also generated outcomes on a continuous scale of 0 to 100 rather than on the five-point ordinal scale more commonly found in radiology. A continuous scale allows for more variability in sensitivity and specificity. Thus, a comparison between continuous scales would be more informative than a comparison between corresponding ordinal scales, because there are no ties in scoring.

Simulation studies examine the behavior of estimators in specific settings. The present simulation study examined plausible scenarios. In radiologic research, test outcomes often are measured by using a five-point ordinal scale, and these tests can be highly accurate, with relatively high specificity. Overall sample sizes often are small as well, including less than 100 subjects. Some important assumptions, however, may have limited the conclusions that can be drawn from the simulation study findings. One important assumption made for these simulations was that the two compared tests had the same underlying ~~sensitivities~~ specificities. Perhaps the strongest assumption made for the simulated data was that when two readers were involved, each had identical ROC curves. In real-life settings, the readers' ROC curves almost certainly will differ. In this context, investigators must determine whether estimating the average AUC

statistic is valuable.

REFERENCES

1. Zerhouni EA, Rutter CM, Hamilton SR, et al. CT and MRI imaging in the staging of colorectal carcinoma: report of the Radiologic Diagnostic Oncology Group II. Radiology 1996; 200:443-451.

2. Baum RA, Rutter CM, Sunshine JH, et al. Multi-center trial to evaluate peripheral vascular magnetic resonance angiography. JAMA 1995; 274:875-880.

3. Song HH. Analysis of correlated ROC areas in diagnostic testing. Biometrics 1997; 53:370-382.

4. Leisenring W, Pepe MS, Longton G. A marginal regression modeling framework for evaluating medical diagnostic tests. Stat Med 1997; 16:1263-1281.

5. Lipsitz LR, Harrington DP. Analyzing correlated binary data using SAS. Comput Biomed Res 1990; 23:268-282.

6. Obuchowski NA. Nonparametric analysis of clustered ROC curve data. Biometrics 1997; 53:567-578.

7. Pepe MS. Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. Biometrics 1998; 54:124-135.

8. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143:29-36.

9. Lee AJ. U-statistics: theory and practice. New York, NY: Dekker, 1990.

10. Bickel PJ, Freedman DA. Some asymptotic theory for the bootstrap. Ann Stat 1981; 9:1196-1217.

11. Arcones MA, Gine E. On the bootstrap of *U* and *V* statistics. Ann Stat 1992; 20:655-674.

12. Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. Crit Rev Diagn Imag 1989; 29:307-335.

Table 1
Average Observed Correlation of Simulated Rating Data (Kendall's τ) ~~Between Tests~~

Correlation Type	Correlation Structure		
	Independence	Compound Symmetry	Disease Dependent
Between tests	0.199	0.478	0.478
Between sites, same disease state	0.164	0.457	0.457
Between sites, different disease state	0.000	0.348	−0.240
Between readers	0.200	0.478	0.478

Note.—Correlation between tests when both tests are assessed by reader 1 with AUC = 0.80, between sites when both are assessed by reader 1 using test A, and between readers when both evaluate the same site with test A.

Table 2
Observed Coverage Rates of 95% Confidence Intervals for AUC Differences

Correlation Structure	Estimator	Sampling Design		
		Small *N*	Large *m*	Large *N*
Independent	Analytic	0.934	0.932	0.946
	1-Reader bootstrap	0.939	0.944	0.946
	2-Reader bootstrap	0.935	0.947	0.949
Equicorrelated	Analytic	0.936	0.934	0.952
	1-Reader bootstrap	0.939	0.946	0.950
	2-Reader bootstrap	0.939	0.947	0.950
Disease-Dependent	Analytic	0.939	0.936	0.950
	1-Reader bootstrap	0.940	0.943	0.948
	2-Reader bootstrap	0.938	0.944	0.950

Note.—Coverage rates are for $AUC_A - AUC_B$, where $AUC_A = 0.8$ and $AUC_B = 0.80$ (5,000 simulations) or $AUC_B = 0.85$ (5,000 simulations). Small *N* simulations generated data from 50 patients with and 50 patients without disease, each of whom was evaluated at four sites during both tests. Large *m* simulations generated data from 100 patients with disease, each of whom was evaluated at 15 sites during both tests. Large *N* simulations generated data from 250 patients with and 250 patients without disease, each of whom was evaluated at four sites during both tests.

Table 3
Observed Coverage Rates of 95% Confidence Intervals for Specificity

Correlation Structure	Small *N*	Large *m*	Large *N*
	Specificity = 0.95		
Independent observations			
Asymptotic normal	0.929	0.936	0.941
Percentile	0.940	0.939	0.944
Equicorrelated (correlation, 0.5)			
Asymptotic normal	0.912	0.923	0.935
Percentile	0.925	0.926	0.949
Dependent on disease state			
Asymptotic normal	0.911	0.912	0.939
Percentile	0.925	0.925	0.949
	Specificity = 0.50		
Independent Observations			
Asymptotic normal	0.942	0.942	0.945
Percentile	0.944	0.949	0.948
Equicorrelated (correlation, 0.5)			
Asymptotic normal	0.941	0.941	0.948
Percentile	0.949	0.945	0.949
Dependent on disease state			
Asymptotic normal	0.941	0.941	0.949
Percentile	0.949	0.941	0.948

Note.—Coverage rates were calculated with data from 10,000 simulations. Small *N* simulations generated data from 50 patients with and 50 patients without disease, each of whom was evaluated at four sites during both tests. Large *m* simulations generated data from 100 patients with disease, each of whom was evaluated at 15 sites during both tests. Large *N* simulations generated data from 250 patients with and 250 patients without disease, each of whom was evaluated at four sites during both tests.

$$\mathrm{AUC} = \frac{\sum_{(i,j)\in D} \sum_{(i',j')\in \bar{D}} \psi(t_{ij}, t_{i'j'})}{\eta_D \eta_{\bar{D}}}$$

Script n, not eta

$$\psi(t_{ij}, t_{i'j'}) = \begin{cases} 1 \text{ if } t_{ij} > t_{i'j'} \\ \frac{1}{2} \text{ if } t_{ij} = t_{i'j'} \\ 0 \text{ if } t_{ij} < t_{i'j'}. \end{cases}$$

Au- symbol missing? Space OK?

$$\frac{(1-p_s)}{(1-p_p\, p_s)} \frac{1}{N}.$$

yes or put a "·" multiplication dot. Either would be accurate

$$\text{sensitivity}_k = \frac{1}{\eta_D} \sum_i \phi_k(\mathbf{t}_i, \mathbf{d}_i)$$

this should be script n not eta

$$\text{specificity}_k = \frac{1}{\eta_{\bar{D}}} \sum_i \{1 - \phi_k[\mathbf{t}_i, (1-d_i)]\}$$

Script n, not eta

Appendix C:
Assessing Mammographers' Accuracy:
A comparison of clinical and test performance

ELSEVIER

Journal of Clinical Epidemiology 53 (2000) 443–450

Journal of Clinical Epidemiology

Assessing mammographers' accuracy: A comparison of clinical and test performance

Carolyn M. Rutter*, Stephen Taplin

Group Health Cooperative of Puget Sound, Center for Health Studies, Seattle, Washington, USA

Received 8 February 1999; received in revised form 24 August 1999; accepted 17 November 1999

Abstract

Direct estimation of mammographers' clinical accuracy requires the ability to capture screening assessments and correctly identify which screened women have breast cancer. This clinical information is often unavailable and when it is available its observational nature can cause analytic problems. Problems with clinical data have led some researchers to evaluate mammographers using a single set of films. Research based on these test film sets implicitly assumes a correspondence between mammographers' accuracy in the test setting and their accuracy in a clinical setting. However, there is no evidence supporting this basic assumption. In this article we use hierarchical models and data from 27 mammographers to directly compare accuracy estimated from clinical practice data to accuracy estimated from a test film set. We found moderate positive correlation [$\hat{\rho} = 0.206$ with 95% credible interval (−0.142–0.488)] between mammographers' overall preponderance to call a mammogram positive. However, we found no evidence of correlation between clinical and test accuracy [$\hat{\rho} = -0.025$ with 95% credible interval (−0.484–0.447)]. This study is limited by the relatively small number of mammographers evaluated, by the somewhat restricted range of observed sensitivities and specificities, and by differences in the types of films evaluated in test and clinical datasets. Nonetheless, these findings raise important questions about how mammographer accuracy should be measured.

Keywords: Sensitivity; Specificity; Hierarchial models; Mammography

1. Introduction

Screening mammography is an effective method of detecting early stage breast cancer. However, the diagnostic value of a mammogram depends on both the technical quality of the film and a mammographer's ability to interpret that film. In the last decade mammographic technology has been relatively stable, allowing researchers to focus on the subjective interpretation of mammograms (e.g., [1,2]).

The Mammography Quality Standards Act recognized the effect of mammographers' interpretations on screening assessments and encouraged medical audits of mammographers' clinical assessments. Evaluating mammographers' performance using clinical assessments is intuitively appealing, because this is "real life" performance. For many researchers, the medical audit is the gold standard measure of performance [3]. However, our ability to draw conclusions about the performance of particular mammographers from these clinical assessments is limited because each mammographer reviews a different set of films. The difficulty of films varies with characteristics of the women evaluated (e.g., breast density), characteristics of lesions (e.g., size), and characteristics of technical film quality (e.g., positioning). Variability in film difficulty results in chance differences among mammographers. Systematic differences in the difficulty of films reviewed can also occur, for example, when mammographers tend to send difficult cases to a particular colleague. Differences in the number of films reviewed also affects comparisons between mammographers through the variability of estimated performance. Because performance estimates based on fewer patients tend to be more variable, and therefore more extreme, comparisons that ignore differences in variability can be misleading. Statistical models have a limited ability to adjust for differences in the films read by each mammographer [4,5].

Estimation of clinical accuracy is further complicated by the influence that clinical assessments have on the probability of detecting breast cancer. When estimating screening accuracy, we focus on the correspondence between a mammographer's clinical interpretation and a woman's true disease state. Because most women only undergo biopsy if a mammographer finds an abnormality, undetected breast cancer cases emerge symptomatically or during a second screening exam. Thus, undetected breast cancer can only be

* Corresponding author. Group Health Cooperative, Center for Health Studies, 1730 Minor Avenue, Suite 1600, Seattle, WA 98101, USA. Tel: 206-287-2190; fax: 206-287-2871.
E-mail address: rutter.c@ghc.org

0895-4356/00/$ – see front matter
PII: S0895-4356(99)00218-8

identified when follow-up information exists. A one year follow-up is generally used, with women classified as disease positive at the time of a screening mammogram if breast cancer is diagnosed within one year [3].

Estimation and comparison of clinical screening performance is also hampered by the relatively low incidence of breast cancer. The one year incidence of invasive breast cancer is approximately 3.5 per 1,000 among American women who are over 49 years old [6]. Low incidence rates make it difficult to precisely estimate a mammographer's rate of cancer detection, since most mammographers will evaluate very few cancers in a single year.

Standardized testing of mammographers is an alternative way to estimate their accuracy. Using standardized film sets removes many of the problems with clinical data. Each mammographer views the same films in the same setting and with the same patient information. Test sets exclude films from women without necessary follow-up information, so that true disease state is known with a high degree of certainty. Test sets can also include more films from women with breast cancer than would be seen in clinical practice, allowing more precise estimation of sensitivity. In summary, use of a test film set controls for film difficulty, film quality and the information presented during film evaluation, offering a relatively simple method of estimating mammographers' accuracy under standardized conditions.

Although estimating accuracy from assessments of standardized film sets avoids many of the problems with clinical data, the artificial conditions introduce other problems. Mammographers know that in the test setting their decisions will not affect patient care. The test itself may be burdensome given time constraints. There is also evidence suggesting that the higher prevalence of disease in test film sets introduces bias. Egglin [7] found that radiologists were more likely to interpret arteriograms as positive for pulmonary emboli when viewed in a higher prevalence film set, regardless of true disease state. When this "context bias" exists, sensitivity increases with increasing prevalence while specificity decreases.

Studies describing mammographer variability based on test film sets (e.g., [1,2]) implicitly assume a strong correlation between mammographers' performance estimated from test sets and mammographers' performance in clinical practice. However, this assumption has never been tested. In this article we directly compare mammographers' clinical and test performance.

2. Data

We analyzed data from 27 mammographers practicing at a large staff model not-for-profit health maintenance organization (HMO). The mammographers included in this study were voluntary participants, though this group essentially included all of the mammographers practicing with the HMO at the time of the study.

Both clinical and test data sets use films from women who remained enrolled in the HMO for at least two years after their index mammogram. Women with breast cancer were identified using the regional Surveillance Epidemiology and End Result registry [8]. Our reference standard for true disease state called a woman "disease positive" at the time of her screening mammogram if either invasive cancer or ductal carcinoma in situ were detected within the following two years. We used a two-year definition because routine follow-up care included mammographic follow-up at either one-year or two-year intervals, depending on a woman's particular risk factors for breast cancer.

2.1. Clinical data

Clinical data used mammographers' final interpretations and recommendations based on mammograms from asymptomatic women screened from 1990 through 1994. Mammographers' interpretations and recommendations have been collected as part of clinical practice for every mammogram evaluated since 1986, using standardized data collection forms. During the time period we examined, mammographer interpretations could be coded as "negative," "inconclusive," or "positive." Final interpretations and recommendations were combined and coded into one of five possible clinical assessments: (1) negative mammogram and recommendation for mammographic follow up at 1 year or later; (2) inconclusive mammogram and recommendation for mammographic follow up at 1 year or later; (3) inconclusive mammogram and recommendation for follow up in less than 1 year (short interval follow-up); (4) inconclusive mammogram and recommendation for biopsy or surgical referral; and (5) positive mammogram.

2.2. Test data

Mammographers were evaluated using test film sets during late 1994 and early 1995. As part of an educational intervention, each mammographer assessed the same set of screening mammograms. Test mammograms were drawn from the population of women screened between 1985 and 1991, using stratified random sampling. Most (92.5%, 111/120) films were selected from the 1990/1991 time period. Films were stratified by the woman's true disease state and the original (clinical) mammographer's assessment. We defined recommendations for short interval follow-up, request for additional work-up, referral to biopsy, and positive mammogram interpretations as positive mammographic assessments, corresponding to clinical assessment categories 3, 4, and 5. Based on each screened woman's true state and dichotomous clinical assessment, we created four strata: (1) true positive (TP) films (positive assessment, breast cancer within one year); (2) false negative (FN) films (negative assessment, breast cancer within one year); (3) true negative (TN) films (negative assessment, no breast cancer); and (4) false positive (FP) films (positive assessment, no breast cancer). From these strata, we randomly selected 23 TP films, 9

FN films, 72 TN films, and 16 FP films. Because of the stratified sampling scheme the test film set was not representative of the mix of films seen in clinical practice: it included an excess of films from women with breast cancer and films that originally lead to incorrect assessments. Out of these 120 films, 7 films (3 TP films and 4 FP films) were excluded from analyses because marks were placed on films during the course of the study. To allow correspondence with the clinical analyses, the reference standard for test films was recalculated, using a 2-year follow-up period. Applying the 2-year follow-up caused one TN film to be recoded as a FN. Within the 113 test mammograms used for analyses, original readers were 67% sensitive and 86% specific. The average age of screened women who contributed films to the test set was 50 years, ranging from 40 to 87 years.

Mammograms were displayed at each participating mammography clinic in a dedicated reading room. Films were displayed in four sets of 30, and each set was displayed for 2 weeks. Mammographers scheduled a time to review films and were given 1 hour to read each set of 30 films. Each "film" included a two-view mammogram, representing a single screening event, and the woman's most recent prior two-view screening mammogram. Prior mammograms were unavailable for 43 women (38%). No additional clinical information was provided, and mammographers were not provided with the disease prevalence in the test set. Mammographers provided one rating for each breast, using standardized data collection forms. The 5 possible screening assessments were: (1) negative or benign; (2) probably benign (short interval follow-up needed); (3) possibly abnormal (additional views needed); (4) suspicious abnormality (biopsy should be considered); and (5) highly suggestive of malignancy. Each mammographer provided data that was at least 98% complete (222/226 ratings) and 15 of the 27 mammographers provided complete data. There were no apparent patterns of missing data between mammographers. These breast-level ratings were recoded as woman-level assessments. If the woman was diagnosed with breast cancer within two years of the mammogram, then the rating given to the breast with disease was used in the analyses. If the woman did not develop cancer in the following two years, then the maximum of the two breast ratings was used.

3. Methods

We are primarily interested in the degree of correlation between mammographers' accuracy measured in a clinical setting and accuracy measured in a test setting. The accuracy measures we focused on are sensitivity and specificity. Sensitivity is the proportion of women with breast cancer who had a positive mammogram assessment. Specificity is the proportion of women without breast cancer who had a negative mammogram assessment.

Calculation of sensitivity and specificity requires definition of a positive assessment. For clinical assessments, we defined ratings 3, 4, and 5 as positive mammograms, corresponding to recommendations for short interval follow-up or biopsy. Unfortunately, test assessments do not completely match clinical assessments. This is partly because clinical assessments were based on final recommendations whereas the test scale included a recommendation for additional views. Clinical data did not include recommendations for additional views because this is an intermediate clinical recommendation, with final recommendations based on these additional views. Given the difference in these two measurement scales, we defined positive outcome in the test set as a recommendation for short interval follow-up, additional views, or biopsy in the test data set, corresponding to ratings 2, 3, 4, or 5. Mammographers' ratings of test films were based on an explicitly ordinal scale that defined a recommendation for additional films (possibly abnormal) as more strongly indicative of disease than a recommendation for short interval follow-up (probably benign).

3.1. Statistical model

We used a hierarchical model to describe mammographers' test and clinical performance measures, and to examine relationships between these measures (see Table 1). Each mammographer contributed data from two 2 × 2 tables, showing the overall agreement between their assessments and womens' disease state.

The model we use accounts for within mammographer variability in estimated sensitivity and specificity by modeling the number of positive assessments each mammographer gave to diseased (y_{ij11}) and not-diseased (y_{ij01}) women with Binomial(n_{ij1},π_{ij1}) and Binomial(n_{ij0},π_{ij0}) distributions. By using the observed sample sizes in Binomial distributions for each mammographer and data set, the model accounts for differences in the amount of data available. The binomial probability of a positive test is based on receiver operating characteristic models [9], and is given by:

$$\pi_{ijk} = \text{logit}^{-1}(\theta_{ij} + \alpha_{ij}D_{ijk}).$$

If D_{ijk} was coded 0 for disease negative films and 1 for disease positive films, then under this model the i^{th} mammographer evaluates the j^{th} data set with specificity equal to $1-\text{logit}^{-1}(\theta_{ij})$ and sensitivity equal to $\text{logit}^{-1}(\theta_{ij} + \alpha_{ij})$. It is simpler to explain the interpretation of θ_{ij} and α_{ii} in terms of false positive rates (equal to 1 − specificity) and true positive rates (equal to the sensitivity). The parameter θ_{ij} cap-

Table 1
Notation used to denote observed counts of films by mammographer (i), data source (j), disease state, and interpretation

	Mammographic interpretation		
	Negative	Positive	
Breast cancer			
No	y_{ij00}	y_{ij01}	n_{ij0}
Yes	y_{ij10}	y_{ij11}	n_{ij1}

Note: Where i = 1, ..., m, indicates mammographer and j = 1,2 indicates the data source (1 = test and 2 = clinical).

tures the i^{th} mammographer's overall tendency to give positive assessments, so that true positive rates increase with increasing false positive rates. The parameter α_{ij} captures the difference between true positive and false positive rates and measures the log-odds ratio of a positive test for films with breast cancer relative to films without breast cancer. As in the ROC context, we call θ_{ij} "cutpoint parameters" and α_{ij} "accuracy parameters."

The parameters θ_{ij} and α_{ij} could be calculated directly from the data. However, they are not estimable when either sensitivity or specificity is 100%, a situation that is more likely when a mammographer evaluates few films. The hierarchical model uses all available information to better estimate these individual parameters. Under the hierarchical model, both cutpoint parameters (θ_{ij}) and accuracy parameters (α_{ij}) are assumed to vary across mammographers and data sources. We assume θ_{ij} and α_{ij} follow a bivariate normal distribution, implemented as:

$$\theta_{i1}|\Theta_1, \sigma_{\theta 1} \sim N(\Theta_1, \sigma^2_{\theta 1})$$

$$\alpha_{i1}|\Lambda_1, \sigma_{\alpha 1} \sim N(\Lambda_1, \sigma^2_{\alpha 1}),$$

where Θ_{i1} and α_{i1} are conditionally independent and

$$\theta_{i2}|\theta_{11}, \theta_{21}, \ldots, \theta_{m1}, \Theta_2, \tau, \sigma_{\theta 2} \sim N\left(\Theta_2 + \tau\left(\theta_{i1} - \frac{1}{m}\Sigma_{i=1}^{m}\theta_{i1}\right), \sigma^2_{\theta 2}\right)$$

$$\alpha_{i2}|\alpha_{11}, \alpha_{21}, \ldots, \alpha_{m1}, \Lambda_2, \lambda, \sigma_{\alpha 2} \sim N\left(\Lambda_2 + \lambda\left(\alpha_{i1} - \frac{1}{m}\Sigma_{i=1}^{m}\alpha_{i1}\right), \sigma^2_{\alpha 2}\right),$$

where Θ_{i2} and α_{i2} are conditionally independent. Thus, the model assumes that within each data set, mammographers' cutpoint and accuracy parameters are (conditionally) independent.

Because the regression model is centered, the expected value of θ_{i2} is Θ_2 and the expected value of α_{i2} is Λ_2. Assuming that θ_{ij} and α_{ij} are normally distributed and linked via a regression model allows fuller use of the available data, resulting in better estimation. Mammographer's cutpoint and accuracy parameters are smoothed toward overall expected values Θ_j and Λ_j, with the degree of smoothing determined by the amount of data each contributes to the model. Estimates for mammographers with less data will tend to be nearer to expected values than estimates for mammographers with more data, while corresponding interval estimates widen to reflect lack of information available for these parameters.

The linear regression models for θ_{i2} and α_{i2} allow different expected performance for the two film sets and build in correlation between cutpoint parameters and correlation between accuracy parameters, with:

$$\text{corr}(\theta_{i1},\theta_{i2}) = \rho_\theta = \frac{\tau\sigma_{\theta 1}}{\sqrt{\tau^2\sigma^2_{\theta 1}\left(\frac{m-1}{m}\right) + \sigma^2_{\theta 2}}}$$

$$\text{corr}(\alpha_{i1},\alpha_{i2}) = \rho_\alpha = \frac{\lambda\sigma_{\alpha 1}}{\sqrt{\lambda^2\sigma^2_{\alpha 1}\left(\frac{m-1}{m}\right) + \sigma^2_{\alpha 2}}}$$

These correlation parameters are more informative than the between dataset correlation of sensitivity of specificity. Correlation in sensitivity and specificity can be driven by mammographers' overall tendency to call a film positive. The correlation parameters ρ_θ and ρ_α separate the overall tendency to call a film positive from the ability to distinguish between films from women with and without breast cancer. Under this model, ρ_θ measures the correlation between cutpoint parameters that are associated with overall preponderance to call a film "positive" while ρ_α measures association between accuracy parameters that are independent of these cutpoint parameters.

The hierarchical model is completed by specifying prior distributions for the remaining unknown parameters. Priors were chosen to cover the range of plausible values of parameters and were selected to be uninformative. We used a Normal(0,10) prior for Θ_1, Θ_2, Λ_1, and Λ_2, and a Normal(0,100) prior for τ and λ. We used an inverse gamma, $\Gamma^{-1}(0.5,2)$, for $\sigma_{\theta 1}$, $\sigma_{\theta 2}$, $\sigma_{\alpha 1}$ and $\sigma_{\alpha 2}$. This prior is diffuse, but does not overweight large values. Quartiles of the $\Gamma^{-1}(0.5,2)$ distribution are 3.03, 8.80, and 39.41. The parameters Θ_1, Θ_2, Λ_1, Λ_2, τ, λ, $\sigma_{\theta 1}$, $\sigma_{\theta 2}$, $\sigma_{\alpha 1}$, and $\sigma_{\alpha 1}$, and $\sigma_{\alpha 2}$ are assumed to be mutually independent.

This model was estimated using the BUGS program [10]. To improve estimation, the disease state indicator D_{ijk} was centered so that $D_{ijk} = 1/2$ for disease positive films and $D_{ijk} = -1/2$ for disease negative films. This transformation does not affect the interpretation of the parameters α_{ijk} and θ_{ijk}. Standard model diagnostics were used to assess convergence of the sampler, as described in the CODA manual [11]. These models resulted in estimated posterior distributions for the model parameters. We present estimated posterior modes and 95% credible intervals based on the 2.5% and 97.5% percentiles. The posterior mode was estimated by the posterior mean for approximately symmetric distributions, and by the posterior median for skewed posterior distributions.

4. Results

There was wide variability in the amount of clinical data available for each mammographer (Table 2). The 27 mammographers clinically evaluated an average of 1890 films during the four-year period (range 232 to 3818), and saw an average of 15 mammograms from women with breast cancer (range 1 to 32). The average clinical prevalence rate across mammographers was 8 cancers per 1,000 mammograms.

Plots of the sensitivity and specificity suggest moderate positive correlation between clinical and test performance. Fig. 1 shows that overall, mammographers tended to be both more sensitive and more specific in clinical practice. The observed

Table 2
Mammographic assessments of 27 mammographers: rate of positive assessments, indicating disease, with the total number of assessments in parentheses

	Test data		Clinical data	
Mammographer	Specificity (N)	Sensitivity (N)	Specificity (N)	Sensitivity (N)
1	0.880 (83)	0.897 (29)	0.922 (1715)	1.000 (14)
2	0.687 (83)	0.833 (30)	0.816 (1492)	1.000 (14)
3	0.687 (83)	0.833 (30)	0.804 (2341)	0.929 (14)
4	0.880 (83)	0.833 (30)	0.823 (2129)	0.933 (15)
5	0.867 (83)	0.800 (30)	0.896 (2818)	0.880 (25)
6	0.756 (82)	0.733 (30)	0.917 (2221)	0.941 (17)
7	0.867 (83)	0.767 (30)	0.965 (1733)	0.684 (19)
8	0.904 (83)	0.700 (30)	0.911 (2045)	0.917 (12)
9	0.867 (83)	0.833 (30)	0.879 (1742)	0.826 (23)
10	0.831 (83)	0.800 (30)	0.832 (1435)	0.833 (12)
11	0.867 (83)	0.800 (30)	0.915 (3299)	0.935 (31)
12	0.831 (83)	0.724 (29)	0.865 (230)	1.000 (2)
13	0.867 (83)	0.833 (30)	0.870 (971)	0.800 (10)
14	0.904 (83)	0.867 (30)	0.877 (675)	0.500 (2)
15	0.783 (83)	0.833 (30)	0.881 (2546)	0.955 (22)
16	0.880 (83)	0.800 (30)	0.930 (441)	1.000 (1)
17	0.855 (83)	0.867 (30)	0.883 (3167)	0.960 (25)
18	0.854 (82)	0.867 (30)	0.822 (1451)	1.000 (11)
19	0.771 (83)	0.833 (30)	0.901 (3786)	0.875 (32)
20	0.904 (83)	0.767 (30)	0.905 (1276)	0.714 (7)
21	0.855 (83)	0.833 (30)	0.908 (3186)	0.800 (25)
22	0.904 (83)	0.733 (30)	0.880 (2585)	0.947 (19)
23	0.855 (83)	0.793 (29)	0.943 (1643)	0.846 (13)
24	0.807 (83)	0.828 (29)	0.913 (1726)	1.000 (10)
25	0.819 (83)	0.767 (30)	0.864 (1151)	1.000 (4)
26	0.892 (83)	0.900 (30)	0.920 (2169)	0.842 (19)
27	0.759 (83)	0.833 (30)	0.867 (663)	0.833 (6)

correlation between clinical and test sensitivities was −0.096; correlation between specificities was 0.446.

The hierarchical model accounts for within mammographer variability in sensitivity and specificity and accounts for differences in the number of films read in clinical practice. The model can be used to better estimate each mammographer's clinical and test-based sensitivity and specificity, and thus to better estimate between dataset correlation in sensitivity and specificity. Model-based estimates of sensitivity and specificity combine information from the entire sample with each mammographer's information. The degree to which estimates differ from observed values reflects the amount of data available, the values of other parameter estimates (i.e., $\hat{\theta}_{i1}$, $\hat{\theta}_{i2}$, $\hat{\alpha}_{i1}$, $\hat{\alpha}_{i2}$, $\hat{\tau}$, and $\hat{\lambda}$) and underlying distributional assumptions. Estimates of clinical specificity were equal to model estimates because these were based on large number of films. In contrast, estimates of clinical sensitivity were more strongly influenced by additional information, especially for mammographers who evaluated very few films. Model-based estimates of between dataset correlation of sensitivity and specificity were similar to observed correlation estimates. Correlation between clinical and test sensitivity was −0.185 with 95% credible interval (−0.269, 0.593). Correlation between clinical and test specificity was 0.408 with 95% credible interval (0.161, 0.616).

We found little evidence of correlation between clinical and test performance parameters (Table 3). Our point estimate of correlation between clinical and test cutpoints was moderate ($\rho_\theta = 0.220$) although the 95% credible interval was broad and covered zero. The estimated probability that $\rho_\theta > 0$ was 89.4%. Our point estimate of the correlation between clinical and test accuracies was near zero ($\rho_\alpha = -0.026$).

We found expected overall differences in test and clinical accuracy. The test film set was constructed to be more difficult than films seen in the usual clinical practice, and as expected the estimated mean clinical accuracy parameter (Λ_2) was greater than the estimated mean test accuracy parameter (Λ_1), indicating that overall readers were more accurate when evaluating clinical data than test data.

Point estimates also demonstrated that mammographers had an overall tendency to give more positive assessments in their clinical practice than in the test setting (mode Θ_2 < mode Θ_1), even though the prevalence of breast cancer was much higher in the test setting.

Estimated between mammographer variability tended to be higher in clinical practice than in the test setting (e.g., $\sigma^2_{\theta 2} > \sigma^2_{\theta 1}$ and $\sigma^2_{\alpha 2} > \sigma^2_{\alpha 1}$), possibly reflecting wider variability in the films each mammographer reads in clinical practice, or the relatively small number of cancer films each mammographer evaluated over the course of four years in clinical practice.

5. Discussion

These results represent a comprehensive comparison of mammographers' assessments in test and clinical settings. The clinical data was based on automated collection of mammographers' interpretations and recommendations. The data systems also allowed two-year follow-up of each woman screened. The test data included a relatively large set of 113 mammograms and included 30 cancers. Finally, our statistical model allowed for differences in the number of films each mammographer assessed during clinical practice.

There was general agreement between observed values and hierarchical model results. Mammographers tended to be less accurate when evaluating the more difficult test film set, and tended to give more positive assessments in their clinical practice. Thus, we found no evidence of context bias as described by Egglin [7]. That is, mammographers did not tend to make more positive assessments in the higher prevalence test film set. However, we cannot conclude from this study that context bias does not exist, because the test context included both a higher disease prevalence and a more difficult set of films.

Model-based estimates of between dataset correlation of sensitivity were stronger than observed correlation, and the estimated between dataset correlation of specificity was statistically different from zero. However, between dataset

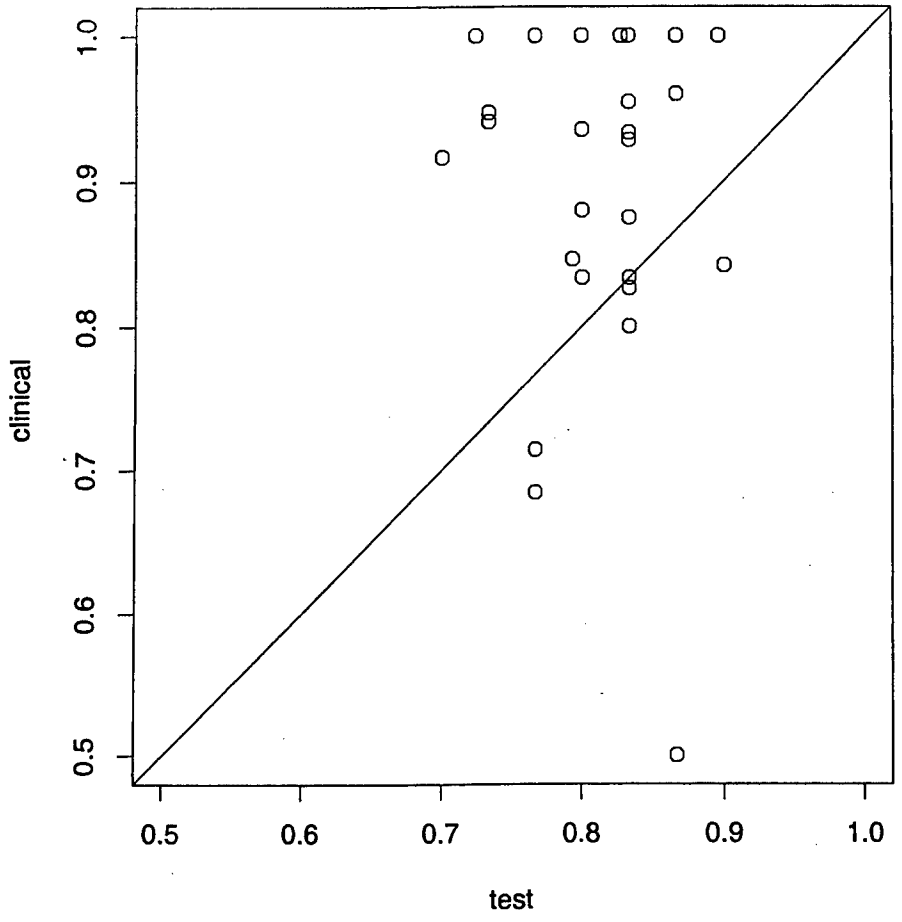


Fig. 1. (A) Sensitivity in clinical practice versus sensitivity in a test setting for 27 mammographers. (B) Specificity in clinical practice versus sensitivity in a test setting for 27 mammographers.

correlation of sensitivity and specificity appears to be driven by correlation in the mammmographers tendency to call tests positive rather than correlation in their accuracy. We found moderate, but not statistically significant, correlation between mammographer's overall preponderance to identify cancer using the two data sources. But there was no apparent correlation between the hierarchical model's accuracy parameters.

We do not believe that the lack of correlation between clinical and test accuracy resulted from differences in outcome scales. The basic assumption that we are testing is that these two measures are correlated because both are mea-

Table 3
Hierarchical model estimates from the posterior distribution

Parameter and description	Estimates	
	Mode	95% Credible Region
Θ_1: expected cutpoint parameter, test data	−0.101	(−0.333, 0.129)
Λ_1: expected accuracy parameter, test data	3.220	(2.900, 3.560)
Θ_2: expected cutpoint parameter, clinical data	0.066	(−0.214, 0.352)
Λ_2: expected accuracy parameter, clinical data	4.360	(3.920, 4.810)
$\sigma^2_{\theta 1}$: between-mammographer variance of cutpoints, test data	0.261	(0.155, 0.489)
$\sigma^2_{\alpha 1}$: between-mammographer variance of accuracy, test data	0.408	(0.212, 0.856)
$\sigma^2_{\theta 2}$: between-mammographer variance of cutpoints, clinical data	0.339	(0.191, 0.672)
$\sigma^2_{\alpha 2}$: between-mammographer variance of accuracy, clinical data	0.505	(0.250, 1.110)
τ: regression coefficient, cutpoint parameters	0.560	(−0.341, 1.530)
λ: regression coefficient, accuracy parameters	−0.048	(−1.020, 0.945)
ρ_θ: correlation between clinical and test cutpoints	0.206	(−0.142, 0.488)
ρ_α: correlation between clinical and test accuracy	−0.025	(−0.484, 0.447)

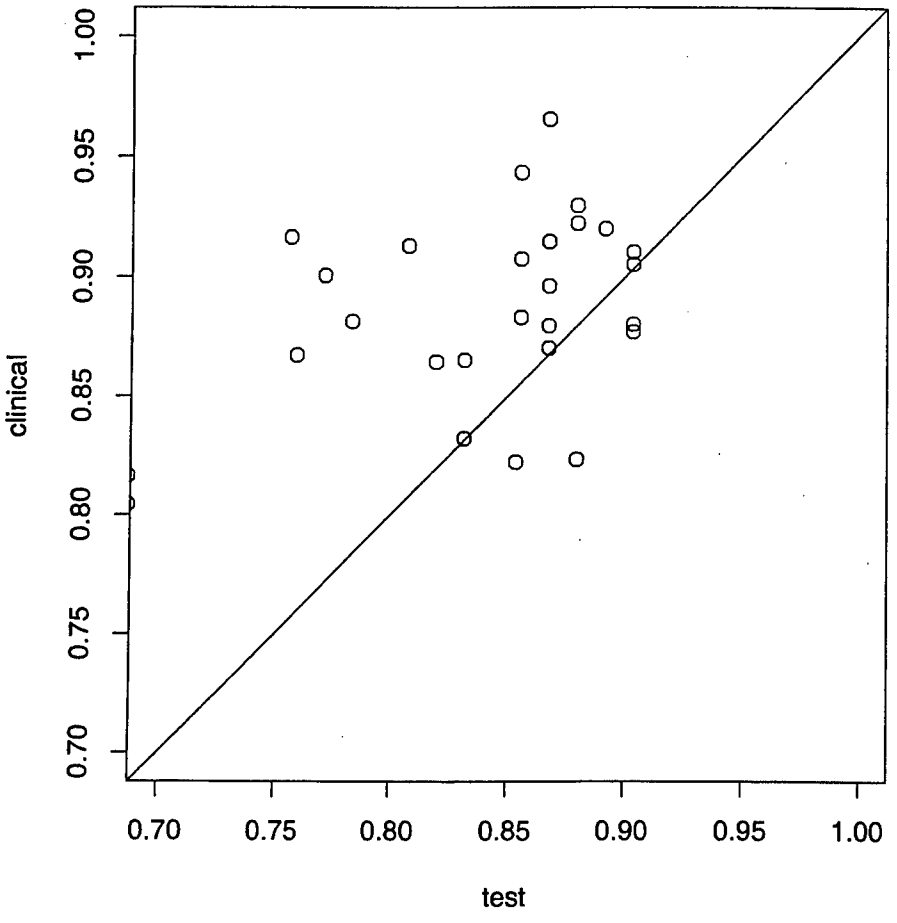


Fig. 1. (Continued)

sures of the same underlying construct, mammographer accuracy. We are not interested in the equality of these two measures; we expect these accuracy estimates to differ because of differences in film difficulty, film quality, and the information provided to mammographers.

We do not believe that the lack of correlation between clinical and test accuracy resulted from dichotomizing the outcome scales. We did not attempt to model the ordinal outcomes directly or via the area under the receiver operating characteristic (ROC) curve because in both clinical and test settings mammographers' maximum false positive rates were relatively low. Because of this, the area under their ROC curves were strongly influenced by false positive rates that were outside of the observed data range especially for clinical data. The sensitivity and specificity pairs we used in analyses contained most of the information available from ROC curves.

Limitations of this study may have prevented us from observing correlation in these data. Our "gold standard" for true disease state was based on a two-year follow-up interval, and misclassification of diseased and not diseased women may have attenuated observed correlation. Our sample of 27 mammographers may have been too small to detect statistically significant correlation, although point estimates suggest there was not clinically relevant correlation in accuracies. Examining mammographers practicing within the same HMO may have reduced the variability of outcomes so that correlation was not observable. Many of the mammographers in this study worked together and discussed difficult cases with each other on a day-to-day basis. Finally, lack of correlation may have resulted from differences in the type of films included in the two data sets. Clinical data included assessments of exams based on imaging studies, such as ultrasound and magnification views. If evaluation of 2 view mammograms requires different skills than evaluation of additional work-up images, then the inclusion of these films in the clinical set could attenuate the estimated correlation between clinical and test accuracy. However, excluding these films would drastically reduce the number of cancer cases included in the clinical set and could bias comparisons by reducing the clinical data set to films that the original reader was able to assess without additional work-up. Because of the limitations of this study, further work is needed to confirm these findings.

The apparent lack of correlation between test and clinical accuracy could be interpreted in at least two ways. One in-

interpretation is that evaluations based on clinical assessments and evaluations based on test film sets are measuring two different kinds of accuracy. Because we are interested in clinical performance, concluding that test-based assessments of accuracy are different from clinical accuracy means either throwing out the test data sets as a reasonable means of mammographer evaluation, or seeking out ways to make test evaluations more comparable to clinical evaluations. A second interpretation is that the apparent lack of correlation between clinical and test performance resulted from differences in the clinical case mix of participating mammographers. Clinical data included assessments based on both standard screening mammograms and screening mammograms that included additional work up, such as ultrasound and magnification views. We do not know how these different types of films were distributed across mammographers, or whether there were any informal systems of referral at the mammography centers. Systematic differences between mammographers could also have been introduced through differences in screened populations, for example, differences in the average age of women screened. Concluding that the clinical data are problematic means either throwing out the clinical data as a means of mammographer evaluation, or seeking out ways to make the clinical evaluations more comparable across mammographers. Unfortunately, our analyses cannot guide our conclusions about clinical and test data, though they caution us against extrapolating results from one setting into another.

Acknowledgments

This study was supported, by grants CA63731 from the National Cancer Institute and BC962461 from the U.S. Department of Defense.

We wish to acknowledge the careful work of Kari Rosvik and Deb Seger who made this study possible, and the many mommorgraphers who gave their time to this study. We want to especially thank Mary Kelly, MD, and Donna White, MD, who provided valuable leadership.

References

[1] Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Archives of Internal Medicine 1996;156:209–13.

[2] Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. New England Journal of Medicine 1994;331:1493–499.

[3] Linver MN, Osuch JR, Brenner RJ, Smith RA. The mammography audit: A primer for the Mammography Quality Standards Act (MQSA). American Journal of Radiology 1995;164:19–25.

[4] Christiansen CL, Morris CN. Improving the statistical approach to health care provider profiling. Annals of Internal Medicine 1997;127: 764–68.

[5] Delong ER, Peterson ED, DeLong DM, Muhlbaire LH, Hackett S, Mark DB. Comparing risk-adjustment methods for provider profiling. Statistics in Medicine 1997;16:2645–664.

[6] Reis LAG, Kosary CL, Hankey BF, Miller BA, Edwards BK (eds). SEER Cancer Statistics Review, 1973–1995. Bethesda, MD: National Cancer Institute; 1998.

[7] Egglin TK, Feinstein AR. Context bias. A problem in diagnostic radiology. Journal of the American Medical Association 1996;276:1752–755.

[8] Miller BA, Reis LAG, Hankey BF. SEER Cancer Statistics Review 1973–1990. NIH Publication No. 93-2789. Bethesda, MD: National Cancer Institute; 1993.

[9] Hanley JA. Receiver operating characteristic (ROC) methodology: The state of the art. Critical Reviews in Diagnostic Imaging 1989;29: 207–35.

[10] Spiegelhalter D, Thomas A, Best N, Gilks W. BUGS 0.6, Bayesian inference Using Gibbs Sampling Manual, MRC Biostatistics Unit, 1997.

[11] Best N, Cowles MK, Vines K. CODA: Convergence Diagnostics and Output Analysis Software for Gibbs sampling output, Version 0.20. Cambridge: MRC Biostatistics Unit; 1995.

Appendix D:
Changes in Breast Density Associated with Initiation, Discontinuation, and Continuing use of Hormone Replacement Therapy (HRT)

unpublished, confidential

in review at JAMA

Changes in Breast Density associated with Initiation, Discontinuation, and Continuing use of Hormone Replacement Therapy (HRT)

Carolyn M. Rutter[1,2], Margaret T. Mandelson[1,3], Mary B. Laya[1,4], Deborah J. Seger[1], Stephen Taplin[1,5]

[1]Group Health Cooperative of Puget Sound

[2]University of Washington, Department of Biostatistics.

[3]University of Washington, Department of Epidemiology.

[4]University of Washington, Department of Medicine, Division of General Internal Medicine.

[5]University of Washington, Department of Family Medicine.

Corresponding author:
Carolyn M. Rutter
Group Health Cooperative of Puget Sound,
Center for Health Studies
1730 Minor Avenue, Suite 1600
Seattle, WA 98101
e-mail: rutter.c@ghc.org
phone: 206.287.2190
fax: 206.287.2871

This study was supported by a career development award to Dr. Rutter from the Department of Defense (DAMD17-97-1-7193) and by cooperative agreement U01 CA63731 and by grant CA71869 (to Dr. Mandelson) from the National Cancer Institute.

word count: 4,297

Abstract

Background: Initiation of hormone replacement therapy (HRT) has been shown to increase breast density (1-5). Several lines of evidence indicate that breast density is strongly related to breast cancer risk (8-11) and that increased density decreases mammographic sensitivity (6).

Objective: To examine the effect of initiation, discontinuation, and continuing use of HRT on breast density using population based data.

Design: Observational cohort study.

Setting: Women enrolled in a large HMO in western Washington state (Group Health Cooperative of Puget Sound).

Participants: 5213 naturally postmenopausal women 40 to 96 years old who had two screening mammograms between 1996 and 1999.

Measurements: HRT use was assessed using automated pharmacy data. Breast density was assessed using clinical radiologists' BIRADS^{TM} ratings.

Results: Women who initiated HRT were more likely than nonusers to show increases in density (OR=3.24, 95% CI (2.47,4.23)), while women who discontinued were more likely show decreases in density (0R=1.92, 95% CI (1.03,3.35)), and women who continued use of HRT were more likely to show both increases in density (OR=1.37, 95% CI (0.89,2.06)) and sustained high density (OR=1.72, 95% CI (1.50,1.98)). Continuing HRT use was more strongly associated with sustained high density among women with high BMI ($p<0.05$).

Conclusions: These results provide strong evidence that breast density changes associated with HRT are dynamic, increasing with initiation and decreasing with discontinuation. Continued HRT use results in persistent changes, particularly among women with high BMI.

Introduction

Several studies have shown that initiation of HRT increases parenchymal breast density (1-5), and there is growing evidence that opposed estrogen has a stronger effect on breast density than unopposed estrogen (4,5). Increases in density induced by HRT can have important consequences. Increased density reduces the accuracy of screening mammography (6) HRT has been directly associated with decreases in both sensitivity and specificity of mammography (7-10), which is likely a result of corresponding increases in density. Studies have also found an association between increased density and increased risk of breast cancer (11-14).

Although the effect of initiating HRT on breast density has been well studied, the effects of discontinuing HRT and continuing HRT have not been systematically examined. In this study, we investigated the relationship between HRT use and density in a population-based cohort of women undergoing at least two screening mammograms. Specifically, we compared changes in breast density across four patterns of HRT use: 1) "Nonusers", who did not use HRT before either mammogram; 2) "Discontinuers", who used HRT before the first mammogram, but not before their second mammogram; 3) "Initiators", who did not use HRT before their first mammogram, but began using HRT before their second mammogram; and 4) "Continuing Users", who used HRT prior to both mammograms. We also explore differential associations between patterns of HRT use and change in breast density by body mass index (BMI).

Methods

Study Sample

Subjects were selected from women enrolled in Group Health Cooperative of Puget Sound (GHC), a health maintenance organization with over 400,000 members in western Washington state. Most mammographic screening at GHC is delivered through the Breast Cancer Screening Program (BCSP), which was established in 1985 (15). The BCSP collects demographic data, health and screening history, and risk-factor information through a self-administered survey mailed to women 40 years of age and older, and generates letters that invite women to begin breast cancer screening and periodically remind them to return for regular screening. During the study period, women were sent screening reminders every 1-2 years, with the reminder interval based on their breast cancer risk factors. GHC physicians may also order screening mammography as part of usual care or for evaluation of symptoms. Data on risk factors and screening examination results, including density assessments, are maintained in a central database.

Women were eligible for our study if they were postmenopausal and had at least two screening exams occurring between January 1996 and December 1998, with the second screening exam occurring at least 11 months, but no more than 25 months, after the first. When women had more than two screening exams during the study period, we chose the pair whose timing was closest to two years apart, corresponding to the most common interval between recommended screenings. Screening consisted of a two-view mammogram and clinical breast examination at dedicated centers within the GHC delivery system.

Women were excluded from our study if they were under 40, had a hysterectomy, had a self-reported history of breast cancer, had a diagnosis of cancer prior to either screening

mammogram, or had undergone breast augmentation. Because we relied on pharmacy data to estimate HRT, we restricted our sample to women who were continuously enrolled in GHC during the year prior to each mammogram.

Measures

Measures of women's HRT use were based on automated pharmacy records. The pharmacy data set captures all prescriptions filled at GHC pharmacies. We defined hormone replacement therapy (HRT) to include estrogens alone and estrogens in combination with a progestin, delivered orally or by patch. Women who filled prescriptions for vaginal rings were excluded from our sample, because unlike other modes of delivery, estrogen delivered via vaginal ring is not associated with higher blood levels of estrogen (16). Women who filled prescriptions for estrogen creams were included in our sample, but because estrogen creams were almost exclusively prescribed for use on an as-needed basis, we did not consider women using creams to be HRT users. We combined pharmacy dose and text instructions to estimate the duration of each prescription and the average dose per day of estrogen and progestin. We estimated the timing of HRT use by assuming that a woman began taking HRT the day after she filled her prescription, with refills extending the duration of HRT use. When a woman filled a prescription for a different HRT drug, or the same drug at a different dosage, within 10 days of an earlier fill, we assumed that her physician had changed either the dose or formulation, effective the day after the new prescription was filled.

We classified HRT use (yes/no) prior to each mammogram using the date of prescription fills and the estimated duration of the prescription. Women classified as HRT users at the time of screening filled a prescription for opposed or unopposed estrogen that lasted for at least 30 days and was estimated to run out no more than six weeks before the screening mammogram. Women classified as non-users at the time of screening had not filled a prescription for estrogen in the prior year, or had filled a prescription that was estimated to run out more than 24 weeks before the screening mammogram.

Our analyses compare four patterns of HRT use: 1) "Nonusers", who did not use HRT before either mammogram; 2) "Discontinuers", who used HRT before the first mammogram, but not before their second mammogram; 3) "Initiators", who did not use HRT before their first mammogram, but began using HRT before their second mammogram; and 4) "Continuing users", who used HRT prior to both mammograms.

In subset analyses we also used women's self-reported current use of HRT in combination with pharmacy data to more stringently classify women using HRT. Self-report data was not used for primary analyses because it does not distinguish as-needed creams from daily preparations, nor does it account for duration or recency of use. Self-reported data are also subject to both reporting errors and missing data. Our refined sub-sample excluded women whose self-reported current HRT use disagreed with pharmacy records.

Breast density was coded on a 4 point scale at the time of each mammogram using American College of Radiology BI-RADS coding (17): 1) almost entirely fat, 2) scattered fibroglandular tissue, 3) heterogeneously dense, and 4) extremely dense. Breast density was coded by clinical radiologists and was captured using an automated reporting system. Radiologists rated density separately for each breast, and the breast with the highest density was used for analysis. To focus on clinically important changes in breast density, we dichotomized density ratings into low (almost entirely fat and scattered

fibroglandular tissue) and high (heterogeneously and extremely dense). For analysis, a change in density was defined as a shift between these dichotomized categories. Therefore, shifts between almost entirely fatty and scattered fibroglandular tissue, and shifts between categories heterogeneously and extremely dense were not considered density changes. "Change" in breast density was coded into four groups: low density (1,2) at both evaluations, decrease in density (3,4)→(1,2), increase in density (1,2)→(3,4), and high density (3,4) at both evaluations.

Cancer outcomes were based on linkage to the Western Washington Surveillance Epidemiology and End Result (SEER) cancer registry. Among women diagnosed with breast cancer within two years of the second exam, we use breast density assessed in the unaffected breast rather than the maximum breast density.

Statistical Analysis

We also examined the association between HRT and density adjusting for two covariates associated with changes in breast density: age at first mammogram and change in body mass index (BMI, kg/m^2). Change in BMI was based on a 5 category measure of change that captured clinically important changes in BMI: 1) lean at both exams (BMI<20), 2) initial BMI between 20 and 25 and a change of less than 1 BMI unit, 3) initial BMI between 20 and 25 and a decrease of at least 1 BMI unit, 4) initial BMI between 20 and 25 and an increase of at least 1 BMI unit, and 5) heavy at both exams (BMI>25). For a woman who is 5'6", a one-unit change in BMI roughly corresponds to a 6 pound weight change. We examined the relationships between HRT and density while controlling for age and change in BMI using three separate logistic regression models to describe the probability of: 1) increased density relative to all other changes, 2) decreased density relative to all other changes, 3) high density at both exams relative to all other changes.

We used stratified logistic regression to examine potential interactions between each covariate (age, BMI) and the effect of HRT change on density change. Age was categorized into 3 groups of approximately equal size (40-49, 50-69, 70+). BMI was grouped into low (<25) versus high (≥25). Differential effects of HRT change were estimated using interaction effects in logistic regression models that adjusted for age and clinically significant BMI change (±1 unit change for women with initial BMI between 20 and 25).

Adjusted relative risks were approximated using a transformation of adjusted odds ratios (18). Adjusted relative risks were used to shed light on differential effects of covariates on odds ratios when there were differences between reference groups.

Results

Among the 6314 women who met initial criteria for inclusion in our sample, 497 (7.9%) had HRT use patterns that did not correspond to one of our four groups. These women had estimated HRT use that ended from 7 to 24 weeks before either mammogram, or had less than 30 days of HRT use prior to a mammogram. One woman who used a vaginal ring prior to her second exam was excluded from our sample. Among the remaining women, 604 (10.4%) were excluded from analyses because of missing density or BMI information (1.1% were missing density and 9.4% were missing BMI). Our final sample included 5212 women with complete density and covariate information.

At the time of the initial mammogram, the average age of women in our sample was 65 years old (sd=9.5, range=40 to 96). There were similar age ranges across the four HRT use patterns, with Nonusers ranging from 43 to 96, Discontinuers from 42 to 80, Initiators from 42 to 91, and Continuing Users from 40 to 92. As shown in Table 1, women who were using HRT at the time of the first mammogram (Discontinuers and Continuing Users) tended to be younger than other groups. Nonusers tended to be older than other groups. Across all groups, approximately one third of women were 60 to 69 years old. However, 46.5% of Nonusers were 70 years or older, while 27.4% of Initiators, 18.9%of Discontinuers and 17.7% of Continuing users were 70 or older.

There were few differences in other risk factors across patterns of HRT use. Our sample was predominately white (92.0%), reflecting the overall racial composition of women enrolled in GHC. 96.1% (4891/5089) had a prior mammogram available to the radiologist at the time of their initial screening mammogram. 15.9% (823/5190) were nulliparous, and 13.8% (594/4299) had their first child after the age of 30. Family history of breast cancer (mother, sister, or daughter) differed across groups, with 24.8% (703/2839) of Nonusers, 20.2% (65/322) of Initiators, 17.1% (18/105) of Discontinuers, and 16.6% (294/1767) of Continuing Users reporting a family history. Overall, there were 42 cancers diagnosed within two years of their second screening exam, and most of these (n=35, 85.4%) were invasive carcinomas (Table 1).

At the time of the initial mammogram, the average BMI of women in our sample was 26.8 (sd=5.5). Differences in average BMI across patterns of HRT use were small, though Discontinuers and Continuing Users tended to be somewhat leaner than other groups (Table 1). About half of the women in this sample (51.7%) had a BMI that was over 25 at both exams. Approximately equal numbers gained one unit or more on the BMI scale (7.6%) or lost one unit or more (7.2%). About one third (29.6) changed less than one BMI unit, and only a few (3.8%) had a BMI that was less than 20 at both exams.

There were no observable differences in HRT dose or drug type across the three groups of women who used HRT (Initiators, Discontinuers, and Continuing Users). Most women who received HRT (93.4%) received a combination of an estrogen and a progestin, with no differences among the three estrogen use groups. Among Continuing Users, only 20 (1.1%) switched between opposed and unopposed estrogen. The most common average daily dose of conjugated estrogen was 0.625mg (62.0% at time 1; 49.8% at time 2). Few women received estrogen doses that were greater than 0.625mg/day (5.1% at time 1; 5.4% at time 2). Some women received doses between 0.15mg/day and 0.50mg/day (15.1% at time 1; 12.1% at time 2). Most women were prescribed Estratab, (86.5% at time 1; 86.8% at time 2); Premarin was the next most commonly prescribed

estrogen (12.3% at time 1; 11.7% at time 2). There was somewhat more variability in the type of estrogen prescribed to Continuing Users than to Discontinuers or Initiators. Among women receiving combination therapy, nearly all were prescribed medroxyprogesterone acetate (98.8% at time 1; 99.0% at time 2), though a few were prescribed norethindrone (11 at time 1; 10 at time 2).

Most (80.6%) women had index screening exams during 1996, and most (64.7%) women had their second mammogram between 21 and 25 months after their first. On average, mammograms were separated by 21 months (sd=3.4). There were no differences in timing of the index mammogram or time between mammograms across patterns of HRT use. There was little loss of information due to use of the maximum density; 99.0% of the women in our sample had equal density ratings in both breasts. Using these maximum ratings, we found expected differences in the distribution of density at the first exam (Table 1). Women using HRT at the time of the first mammogram (Discontinuers and Continuing Users) tended to have higher density at the initial screening than women who were not using HRT at the time of their first mammogram (Initiators and Nonusers).

Table 2 shows the overall relationship between patterns of HRT use and change in breast density. Relative to the Nonuser group, Initiators were more likely to have an increase in breast density (28.4% versus 11.6%), Discontinuers were more likely to have a decrease in breast density (12.6% versus 6.5%), and Continuing Users were more likely to have high breast density at both exams (41.9% versus 25.7%). We reanalyzed the data and found very similar results for 3 subsets of women with more stringent definitions of HRT use: excluding 302 women who were not using HRT, but were using estrogen cream; excluding 173 women using unopposed estrogen; and excluding 266 women whose self-reported HRT use differed from pharmacy records.

Table 3 shows estimated associations between pattern of HRT use and change in breast density based on logistic regression models that adjusted for age at initial exam and change in BMI between the two mammographic screening exams. Because few women had BMI below 20 at both exams, these women were grouped with women who had a less than one-unit change in BMI between evaluations. Initiation of HRT was significantly associated with higher odds of an increase in density (OR=3.24, 95% CI (2.47,4.23)) and lower odds of a decrease in density (OR=0.47, 95% CI (0.24,0.84)). Discontinuation of HRT was significantly associated with higher odds of a decrease in density (OR=1.92, 95% CI (1.03,3.35)). Continuing HRT use was significantly associated with higher odds of an increase in density (OR=1.38, 95% CI (1.15,1.66)) and higher odds of high density at both exams (OR=1.72, 95% CI (1.50,1.98)). Analyses that excluded women whose self-reported HRT use differed from pharmacy records were virtually identical. Analyses that stratified by age (40-49, 50-69, 70+) showed similar results across age groups.

The relationship between pattern of HRT use and change in breast density varied across subgroups of women stratified by BMI (<25 versus ≥ 25) at first mammogram (Table 4). Initiation of HRT was significantly associated with increases in density for both groups, with a somewhat stronger (though not statistically different) effect among women with low BMI. Among women with high BMI, initiation of HRT use was also associated with high density at both mammograms (interaction, $p<0.05$). When interactions with baseline BMI were included, discontinuation of HRT was not significantly associated with changes in breast density. Continuing HRT use was significantly associated with both

increased density and high density at both exams for women in each BMI groups. The association between continuing HRT use and high density at both exams was significantly stronger among women with high BMI (interaction, $p<0.05$).

Among Nonusers, low BMI (<25) was associated with higher breast density. For example, 39.9% of Nonusers with low BMI had low density at both mammograms, versus 68.0% of Nonusers with high BMI. Similarly, 40.1% of Nonusers with low BMI had high density at both mammograms, versus 15.2% of Nonusers with high BMI. Table 5 shows adjusted relative risks for the outcome most strongly influenced by differences in the reference groups, high density at both exams. Based on adjusted relative risk estimates, Initiators were less likely to have high density at both exams than Nonusers, with a somewhat stronger effect among women with low BMI. Discontinuers with low BMI were more likely to have high density at both exams, while Discontinuers with high BMI were less likely to have high density at both exams. Finally, continuing users were more likely to have high density at both exams, particularly among women with high BMI.

Discussion

This study has several strengths that distinguish it from earlier research. Automated pharmacy data allowed us to measure HRT use across a large group of mammographically screened women whose breast density was routinely recorded. We believe that this is the largest study to date of HRT use and breast density changes. This is the only study to simultaneously examine HRT initiation, discontinuation, and continuing use relative to women not using HRT. We focused on clinically significant changes in breast density, and distinguished women with fatty breasts at both exams from women with dense breasts at both exams. This is also the first published study to explore changes in density adjusting for co-occurring changes in BMI, and the first to examine differential effects of HRT on density for women with high versus low BMI.

Our analyses provide important new information about women who discontinue HRT and women who are continuing HRT users. Discontinuation of HRT was associated with subsequent decreases in density, and increases in breast density were sustained by continued HRT use. We also found that initiation of HRT was associated with increases in parenchymal breast density. These results provide strong evidence that breast density changes associated with HRT are dynamic, increasing with initiation and decreasing with discontinuation.

We found some evidence of differential effects of change in HRT for women with low versus high body mass index (BMI). Continuing HRT users who had high BMI were at greater risk of consistently high mammographic density at both exams than continuing users with low BMI. Recent studies based on retrospective self-report of HRT use have associated HRT use with increased risk of breast cancer primarily among women with low BMI (19,20). Our findings raise new concerns for women with a higher body mass index. Increased density is associated with decreased mammographic accuracy and increased cancer risk. (6) Because postmenopausal women with high BMI are at a higher risk of breast cancer than women with low BMI (21-23), sustained increases in breast density due to HRT use could have particularly deleterious effects among women with high BMI, potentially increasing their breast cancer risk and likely decreasing the accuracy of their screening mammograms.

Our analyses confirmed findings from previous studies demonstrating an association between initiation of HRT use and increases in parenchymal breast density (1-5). Like these earlier studies, we examined changes in breast density among women who began using HRT. Studies that have failed to find a difference compared the parenchymal patterns of women using HRT to patterns of women not using HRT rather than examining within woman changes (24,25).

We adjusted analyses for age and change in BMI because these covariates are known to be related to change in density over time. Previous studies examining the effect of HRT initiation on breast density have not consistently adjusted for age, and none have been able to adjust for change in BMI. Laya and colleagues allowed adjustment for a variety of covariates (including age, weight and height) via stepwise regression. Greendale and colleagues adjusted for age, uterus status, baseline density, cigarette smoking, and alcohol use, but not for baseline BMI or parity because they found that neither affected change in density. Greendale and colleagues did not find evidence of differential effects of HRT for various subgroups of patients, although they did not provide results for baseline BMI. We found differential effects of HRT on sustained high density by BMI. Because Greendale and colleagues focused on increases in density, they excluded women who had high density at baseline, and thus could not have observed this effect.

We measured density using the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS), with assessments made by a variety of clinical radiologists. Studies finding increases in density associated with initiation of HRT have used a variety of measures, though all relied on expert readers. Greendale and colleagues (4) used BI-RADS coding. Several earlier papers were based on the Wolfe classification scheme (2-3,5) and one study (1) used a simple measure of "Dense", "Heterogenous", and "Fatty". These studies consistently demonstrated an association between initiation of HRT and increased breast density across a variety of density measures. Our study is consistent with previous research, showing clear associations between initiation of HRT and increased breast density as measured in clinical practice.

We were able to examine changes in HRT use and breast density changes in a relatively large group of women because of automated data collection. Previous studies of initiators have also used detailed HRT dose information, though the source is sometimes unclear. Two studies appear to rely on a combination of medical records and self report for HRT dose information (1,2,5), while others used information from randomization within in a treatment study. (3,4) Automated pharmacy data allowed broad capture of information, but are also limited. HRT use was based on pharmacy fills, so some women who filled prescriptions but did not subsequently take HRT may have been misclassified. In addition, some women who filled HRT prescriptions at outside pharmacy facilities may have been misclassified as nonusers. Timing information is especially sensitive, since we cannot capture when a woman begins to take her prescription. Because of these limitations, we used relatively stringent requirements for categorization into use and non-use groups, resulting in the exclusion of women who had intermediate use patterns. We also examined a subset of women whose self-reported current HRT use was consistent with pharmacy estimated HRT use and found virtually identical results. Overall, only 5.1% of women self-reported HRT use that differed from pharmacy records. Thus, there is little evidence of missclassification of HRT use and such misclassification would result in attenuation of the effect of HRT use on density change.

Like all previous studies of HRT use, the current study is subject to selection bias. This is an inherent problem in studies of HRT use. Women chose whether and when to initiate, continue or discontinue HRT use, and these choices may be related to covariates that in turn bias these analyses. Randomized trials that include a placebo control (4) come closest to controlling for potential selection bias, since all women, including the placebo group, were willing to initiate HRT. Although a recent study of HRT users suggests that there are fewer differences between users and nonusers than previously expected (26), our results must be viewed in light of their observational nature.

In addition, these data did not allow us to address several important factors that may influence the effect of HRT on breast density. We were unable to examine the effects of opposed versus unopposed estrogen and the effects of type of drug prescribed, because there was not enough variability in these factors within our sample, reflecting the selection of naturally postmenopausal (i.e., non-hysterectomized) women. In addition, automated pharmacy data did not allow us to distinguish between cyclical and combination estrogen, and progestin. We also lacked information about women's overall duration of use.

This study shows strong associations between patterns of HRT use and changes in breast density. Our findings suggest that in some women HRT increases breast density but that increases are potentially reversible with cessation of HRT. This result has important implications for breast cancer screening. Increased density adversely affects the accuracy of screening mammography and is a strong, if not the strongest, risk factor for cancer missed at screening (6). HRT itself is associated with decreases in both the sensitivity and specificity of mammographic (7-10). Observed decreases in mammographic accuracy among women using HRT are a likely result of corresponding increases in density.

Previous studies have associated HRT with increased risk of breast cancer, with stronger associations for estrogen combined with progestin (opposed estrogen) than unopposed estrogen (19,20,27-29). Estrogen increases normal breast cell proliferation (30,31), and this increased cell proliferation may be a pathway to both increased breast density and increased risk of breast cancer. Thus, breast density may be an important intermediate outcome on the pathway between HRT and breast cancer. We found that density changes associated with HRT varied across women. These results suggest that density changes associated with HRT use, particularly increases, may be a marker of increased susceptibility to estrogen, and possibly increased risk for breast cancer. Exploration of biological processes related to differential effects of HRT on breast density could illuminate underlying processes related to breast cancer risk and differential effects of HRT on breast cancer risk.

References

1. Stomper PC, Van Voorhis BJ, Ravnikar VA, Meyer JE. Mammographic changes associated with postmenopausal hormone replacement therapy: a longitudinal study. *Radiolgoy.* 1990;174:487-90.
2. Kaufman Z, Garstin WI, Hays R, et al. The mammographic parencymal patterns of women on hormonal replacement therapy. *Clinical Radiology.* 1991;43:389-92.
3. Laya MB, Gallagher JC, Schreiman JS, et al. Effect of postmenopausal hormone replacement therapy on mammographic density and parenchymal pattern. *Radiology.* 1995;196: 433-7.
4. Greendale GA, Reboussin BA, Sie A, et al. Effects of Estrogen and Estrogen-Progestin on Mammographic Parenchymal Density. *Annals of Internal Medicine.* 1999;130: 262-269.
5. Lundstrom E, Wilezek B, von Palffy Z, et al. Mammographic breast density during hormone replacement therapy: Differences according to treatment. *American Journal of Obstetrics and Gynecology.* 1999;18:348-352.
6. Mandelson MT, Oestreicher N, Porter PL, Taplin SH, White E. Breast density as a predictor of mammographic detection: Comparison of interval- and screen-detected cancers. *JNCI. in press.*
7. Laya MB, Larson EB, Taplin SH, White E. Effect of estrogen replacement therapy on specificity and sensitivity of screening mammography. *JNCI.* 1996;88:643-649.
8. Litherland JC, Stallard S, Hole D, Cordiner C. The effect of hormone replacement therapy on the sensitivity of screening mammograms. *Clin Radiol.* 1999;54:285-8.
9. Seradour B, Esteve J, Heid P, Jacquemier J. Hormone replacement therapy and screening mammography: analysis of the results in the Bouches du Rhone programme. *J Med Screen.* 1999;6:99-102.
10. Kavanagh AM, Mitchell H, Giles GG. Hormone replacement therapy and accuracy of mammographic screening. *Lancet.* 2000;355:270-274.
11. Saftlas AF, Szklo M. Mammographic parenchymal patterns and breast cancer risk. *Epidemiol Review.* 1987;9:146-74.
12. Saftlas AF, Hoover RN, Brinton LA, et al. Mammographic densities and risk of breast cancer. *Cancer.* 1991;67:2833-8.
13. Warner E, Lockwood G, Trichler D, Boyd NF. The risk of breast cancer associated with mammographic parenchymal patterns: a meta-analysis of the published literature to examine the effect of method of classification. *Cancer Detect Prev.* 1991;16:67-72.
14. Oza AM, Boyd NF. Mammographic parenchymal patterns: a marker for breast cancer risk. *Epidemiol Review.* 1993;15:196-208.
15. Taplin SH, Mandelson MT, Anderman C, White E, Thompson RS, Timlin , et al. Mammography diffusion and trends in late-stage breast cancer: evaluating outcomes in a population. *Cancer Epidemiol Biomark Prev.* 1997;6:625-631.
16. Gabrielsson J, Wallenbeck I, Birgerson L. Pharmacokinetic data on estradiol in light of the estring concept. Estradiol and estring pharmacokinetics. *Acta Obstet Gynecol Scand, Suppl.* with discussion. 1996;163:26-34.
17. Bassett LW, Feig SA, Jackson VP, Kopans DB, Linver MN, Sickles EA, et al. American College of Radiology ACR Breast Imaing Reporting and Data System BI-RADS™ . Third Edition. Reston (VA): American College of Radiology; 1998.
18. Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. *J Clin Epidemiol.* 1994;84:881-889.
19. Schairer C, Lubin J, Troisi R, et al. Menopausal estrogen and estrogen-progestin replacement therapy and breast cancer risk. *JAMA.* 2000;283: 485-535.

20. Magnusson C, Baron JA, Correia N, Bergstrom R. Adami HO, Persson I. Breast-cancer risk following long-term oestrogen- and oestrogen-progestin replacement therapy. *Int J Cancer.* 1999;81:339-44.
21. Li CI, Stanford JL, Daling JR. Anthropomorphic variables in relation to risk of breast cancer in middle-aged women. *Int J Epidemiol.* 2000;29:208-13.
22. Hirose K, Tajima K, Hamajima N, et al. Effect of body size on breast-cancer risk among Japanese women. *Int J Cancer.* 1999;80:349-55.
23. La Vecchia C, Negri E, Franceschi S, et al. Body mass index and post menopausal breast cancer: an age-specific analysis. *Br J Cancer.* 1997;75:441-4.
24. Bland KI, Buchanan JB, Weisberg BF, et al. The effects of exogenous estrogen replacement therapy of the breast: breast cancer risk and mammographic parenchymal pattern. *Cancer.* 1980;45:3027-33.
25. Berkowitz JE, Gatewood OM, Goldblum LE, Gayler BW. Hormonal replacement therapy: mammographic manifestations. *Radiology.* 1990;174:199-201.
26. Buist DSM, LaCroix AZ, Newton KM, Keenan NL. Are long-term hormone replacement therapy users different from short-term and never users? *A J Epidemiol.* 1999;149:275-81.
27. Colditz GA, Hankinson SE, Hunter DJ, et al. The use of estrogens and progestins and the risk of breast cancer in postmenopausal women. *N Engl J Med.* 1995;332:1589-93
28. Ross RK, Paganini-Hill A, Wan PC, Pike MC. Effect of hormone replacement therapy on breast cancer risk: estrogen versus estrogen plus progestin. *J Natl Cancer Inst.* 1999;92:328-332.
29. The Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormone replacement therapy: collaborative reanalaysis of data from 51 epidemiological studies of 52705 women with breast cancer and 108411 women without breast cancer. *The Lancet.* 1997;350:1047-59.
30. Laidlaw IJ, Clarke RB, Howell A. et al. The proliferation of normal human breast tissue implanted into athymic nude mice is stimulated by estrogen but not progesterone. *Endocrinology.* 1995;136:164-171.
31. Raafat AM, Hofseth LJ, Li S, Bennett JM, Haslam SZ . A mouse model to study the effects of hormone replacement therapy on normal mammary gland during menopause: enhanced proliferative response to estrogen in late postmenopausal mice. Endocrinology. 1999;140:2570-2580.

Table 1: Baseline Characteristics by Patterns of HRT use.

	Nonusers at both times	*Initiated use before second exam*	*Discontinued use before second exam*	*Continuing use at both exams*	*Overall*
	n=2942 (56.4%)	*n=335 (6.4%)*	*n=111 (2.1%)*	*n=1824 (35.0%)*	*n=5213*
Age at first exam, mean (sd)	67.8 (9.3)	63.3 (9.4)	61.2 (8.4)	61.2 (8.3)	65.1 (9.5)
Body Mass Index at first exam, mean (sd)	27.0 (5.6)	27.4 (6.5)	26.5 (5.0)	26.4 (5.2)	26.8 (5.5)
Breast density at first exam					
1) almost entirely fat	18.5%	15.5%	9.9%	6.8%	14.0%
2) scattered fibroglandular tissue	49.4%	53.1%	40.5%	44.3%	47.6%
3) heterogenously dense	27.4%	27.2%	42.3%	38.2%	31.5%
4) extremely dense	4.8%	4.2%	7.2%	10.7%	6.9%
Cancer diagnosis within one year of second exam					
overall	0.7% (n=24)	0.6% (n=2)	0	0.9% (n=16)	0.8% (n=42)
in-situ	0.2% (n= 5)	0	0	0.1% (n= 2)	0.1% (n= 7)
invasive	0.6% (n=19)	0.6% (n=2)	0	0.8% (n=14)	0.7% (n=35)

Table 2: Change in Breast Density by Patterns of HRT use

HRT use:	*Low density, both exams* *n=2509*	*Decrease at 2nd exam* *n=345*	*Increase at 2nd exam* *n=705*	*High density, both exams* *n=1654*	*OVERALL* *n (%)*
Nonusers at both times	56.2%	6.5%	11.6%	25.7%	2942 (56.4%)
Initiated use before second exam	40.3%	3.3%	28.4%	28.1%	335 (6.4%)
Discontinued use before second exam	39.6%	12.6%	10.8%	36.9%	111 (2.1%)
Continuing use at both exams	37.1%	7.0%	14.0%	41.9%	1824 (35.0%)

Table 3: Adjusted associations between patterns of HRT use and changes in breast density.

	Decrease in density at second exam, relative to no decrease		*Increase in density at second exam, relative to no increase*		*High density at both exams, relative to not high density at both exams*	
HRT use:	*OR*	*95% CI*	*OR*	*95% CI*	*OR*	*95% CI*
Initiated use before second exam	0.47	(0.24,0.84)	3.24	(2.47,4.23)	0.98	(0.75,1.27)
Discontinued use before second exam	1.92	(1.03,3.35)	1.02	(0.52,1.81)	1.37	(0.89,2.06)
Continuing use at both exams	1.01	(0.79,1.29)	1.38	(1.15,1.66)	1.72	(1.50,1.98)

Each odds ratio (OR) was estimated relative to any other density change at second exam. Each group of women with some HRT use (Initiators, Discontinuers, and Continuing Users) was compared to women with no HRT use. Odds ratios are adjusted age and change in body mass index.

Table 4: Adjusted associations between patterns of HRT use and changes in breast density by BMI strata

HRT use:	*BMI strata*	*Decrease in density at second exam, relative to no decrease*			*Increase in density at second exam, relative to no increase*			*High density at both exams, relative to not high density at both exams*		
		n+	*OR*	*95% CI*	*n+*	*OR*	*95% CI*	*n+*	*OR*	*95% CI*
Initiated use before second exam	<25	7	0.56	(0.23,1.14)	49	3.96	(2.68,5.82)	53	0.74	(0.52,1.06)
	≥25	4	0.37	(0.13,1.02)	46	2.72	(1.88,3.94)	41	1.34*	(0.92,1.94)
Discontinued use before second exam	<25	5	1.14	(0.38,2.69)	4	0.69	(0.21,1.74)	29	1.57	(0.89,2.79)
	≥25	9	3.04	(1.44,6.39)	8	1.33	(0.62,2.86)	12	1.21	(0.63,2.33)
Continuing use at both exams	<25	70	0.95	(0.68,1.32)	118	1.30	(1.00,1.70)	461	1.44	(1.20,1.73)
	≥25	58	1.08	(0.76,1.53)	138	1.45	(1.14,1.85)	303	2.13*	(1.75,2.59)

Odds ratios are adjusted for age at the time of the initial mammogram and clinically significant change in BMI (initial BMI between 20 and 25 with a change of at least one BMI unit). Estimates are based on logistic regression that includes interactions between BMI strata and HRT use pattern ((Initiators, Discontinuers, and Continuing Users). Overall, 2298 women (44.1%) had BMI below 25 at baseline and 2915 (55.9%) had BMI greater than or equal to 25.

+ number with outcome.

* Significant different odds for high and low BMI groups at the 0.05 level.

Table 5: Adjusted associations between patterns of HRT use and changes in breast density by BMI strata

HRT use:	*BMI strata*	*High density at both exams, relative to not high density at both exams*	
		RR	*95% CI*
Initiated use before second exam	<25	0.53	(0.36,0.75)
	≥25	0.81	(0.55,1.17)
Discontinued use before second exam	<25	1.11	(0.63,1.97)
	≥25	0.73	(0.38,1.40)
Continuing use at both exams	<25	1.02	(0.85,1.22)
	≥25	1.28	(1.06,1.56)

Relative risks are estimated from baseline risk and adjusted odds ratios given in Table 4. Relative risks adjusted for age at the time of the initial mammogram and clinically significant change in BMI (initial BMI between 20 and 25 with a change of at least one BMI unit).

DEPARTMENT OF THE ARMY
US ARMY MEDICAL RESEARCH AND MATERIEL COMMAND
504 SCOTT STREET
FORT DETRICK, MARYLAND 21702-5012

REPLY TO
ATTENTION OF:

MCMR-RMI-S (70-1y) 8 Jan 2003

MEMORANDUM FOR Administrator, Defense Technical Information Center (DTIC-OCA), 8725 John J. Kingman Road, Fort Belvoir, VA 22060-6218

SUBJECT: Request Change in Distribution Statement

1. The U.S. Army Medical Research and Materiel Command has reexamined the need for the limitation assigned to the enclosed. Request the limited distribution statement for the enclosed be changed to "Approved for public release; distribution unlimited." These reports should be released to the National Technical Information Service.

2. Point of contact for this request is Ms. Judy Pawlus at DSN 343-7322 or by e-mail at judy.pawlus@det.amedd.army.mil.

FOR THE COMMANDER:

Encl

PHYLLIS M. RINEHART
Deputy Chief of Staff for
Information Management

ADB265840
ADB279138
ADB264578
ADB281679
ADB281645
ADB261128
ADB261339
ADB273096
ADB281681
ADB259637
ADB256645
ADB262441
ADB281674
ADB281771
ADB281612
ADB266633
ADB251763
ADB281601
ADB258874
ADB281773
ADB281660
ADB259064
ADB266141
ADB281664
ADB258830
ADB266029
ADB281668
ADB259834
ADB266075
ADB281661
ADB282069
ADB265386
ADB282057
ADB258251
ADB264541
ADB241630
ADB281924
ADB281663
ADB281659